ETL Project.

Team Members

Aline Jaimes

Elizabeth Shaw

David Ayankoya

**FINAL REPORT**

# **Open Data for the City of Houston**

## **MOTIVATION.**

The open data website of the City of Houston has listed approximately 250 datasets available to the public. Most of these datasets are outdated our URLs are disabled. Our goal is to create a database of selected updated data sets for the city of Houston

## **METHODS.**

Each member of our team selected a dataset from different websites:

1.David: Greater Houston Area public schools

2.Aline: Texas Commission Environmental Quality dump sites

3.Elizabeth: Texas Commission Environmental Quality dry cleaning waste disposal sites

1.<http://data.houstontx.gov/dataset/public-schools-in-the-houston-region/resource/04da3864-8f59-4b0f-8611-1542b1584d6f>

2. <https://www17.tceq.texas.gov/tamis/index.cfm?fuseaction=report.site_list>

3.[https://www.tceq.texas.gov/agency/data/lookup-data/drycleaners-data-records](https://www.tceq.texas.gov/agency/data/lookup-data/drycleaners-data-records.html)

Because most of the data sources included state/national wide information, the data cleaning process consisted of dropping columns of data that did not correspond to the city of Houston. We explored both methods to export data into a database MongoDB or SQL. Special attention is paid to the county columns in the datasets.

### CLEANING

Dry Cleaning Data:

1. Open csv in pandas.
2. Select only columns relevant for comparison
3. Examine data types. Change zip code object to int.
4. Sort [COUNTY] column so that only relevant Houston Metro counties are present. Drop other counties.
5. Merge sorted county tables into single dataframe.
6. Export to new cleaned data csv

TECQ Data:

1. Scraped data table of monitoring station in the state of Texas
2. Converted to data frame with Pandas
3. Exported the data to a csv
4. Loaded into a SQL database

Public Schools Data:

1. Downloaded CSV file from
2. Pulled into pandas and dropped not relevant columns
3. Renamed columns and kept counties
4. Exported clean data to csv
5. Loaded DataFrame into SQL DataBase

## **RESULTS**

Our final results are contained in a repository called HTownPy.

Codes:

1.TECQ\_DataV2 - Aline

2. School\_data.ipynb - David

3. Dry\_Cleaning.ipynb - Elizabeth