**ETL PROJECT FINAL REPORT**

**EVA and SULEYMAN**

1. **Objectives:**

The main purpose of the project was testing and improving our ETL (Extract, Transform, Load) skills through the ELT project. Keeping that in mind, Eva and I developed an ELT project which is NSPC (Number of School and Population Comparison). In this project we targeted to find out how much population are there per school in Harris County. With that, we needed the two types of data:

* Population of Harris County’s regions, areas, cities etc.
* Schools in these regions, areas, cities, etc.

We faced two important challenges about the data we need. First one is how can we divide the Harris County based on population and schools because it is a huge county with million population and more than 1300 schools at all levels. Second, what was the common points(columns) must be. Because we needed to merge our two datasets based on this common column having the same data in it for the both datasets. Finally, we decided to find and combine the data based on their zip codes and extracted the data accordingly.

1. **Extraction**

We extracted a dataset on Houston Population from Census.gov. The link of our API request was

<https://api.census.gov/data/2017/acs/acs1/groups.html>. Our API code came from a config.py file to extract the data.

The dataset was retrieved by population data based on the zip codes in Harris County and loaded into a Pandas Data Frame. In this dataset, there were two columns: zip codes and 2017 population of these 132 zip codes. Because we wanted to extract only these two data from the data base that we planned to use. After extraction we saved the data as .csv.

The dataset for the “public schools in Houston” was downloaded from <https://data.world>.

This dataset included “objectid\_12”, “objectid\_1”, “objectid”, “distname”, “campus”, “campusname”, “streetmail”, “city\_mail”, “street”, “city”, “zip”, “county”, “cntyname”, “region”, “graderange”, “gradegrp”, “verify”, “instr\_type”, “street\_m\_1”, “adjust”, and shape. This data set was already in csv format and we also saved this file in our project folder.

1. **Transformation**

The population data was cleaned by resetting the index, sorting the zip codes as ascending, dropping the extra index columns, converted and saved the file in csv format.

The necessary columns from the school dataset was selected. These column names included “zip codes”, “districtname”, “campusname” and “graderange” (school type). The columns were renamed and sorted by zip codes.

The data in both population and public school in Houston data were retrieved by common zip codes. And the irrelevant columns were dropped. And both datasets were merged on zip codes. What we basically did here is grouping by the zip codes (because many schools placed in the same zip codes) and count the schools in each zip codes. Our cleaned school data included two columns: grouped zip codes and the number of schools for each zip codes.

To have the final cleaned table we merged the clean population data with the cleaned school data on their zip codes. Now we had the zip codes, their 2017 population and the number of schools in each zip code. There was only one more step to reach our target result which is the number of populations per school in Harris county. So, we calculated the number of populations per each zip codes in the county we divided the total population by the total school number in each zip codes to find it.

As a final shot, we wanted to see how many people there are per school in the county and we divided the total population by the total school number in the Harris county and we found it there are almost 4424 people per school in Harris County.

1. **Loading**

With our data, a PostgreSQL database was created, and connection was established to the database via Jupiter Notebook by using out login name and password. We created an engine and a data base, and placed our tables in this database.

We stored the following tables:

* census\_sorted,
* cleaner, and
* merged.

Eva & Suleyman