Data Preparation Phase Team B

Data Collection and Cleaning

Data was collected from the shared drive, we were told to work with only files that begin from 2016 Fall Semester to the end of 2024 Spring Semester. These reports are also made for several universities per year range. Totalling in around 220 files.

We decided to split the merging into several tasks for everybody at first because a large enough part of the data was not under the same format, so we had to get that fixed first. Every member took an equal part of data to fix, which also helped us do it quicker. After fixing and merging the files in their group, we merged all groups into 1. The dataset contains all of the student addresses, given they lived off-campus, throughout the years since the 2016th Fall Semester. Final file contains arround 290 thousand entries.

Then, we switched to cleaning, we decided to split the cleaning into two parts and assigning each part to a different member of a team, whilst the other group inspected other datasets in the project description to possibly find a connection between 2.

The first part of cleaning is to fix the addresses, the issue was that, the format for address was prepicked for most of them, which was `street_number, street_name, street_suffix, unit_number`, but some were under format where the whole address was written in a span of 1 column, rather than 4. So we had to split the addresses, splitting by spaces didn't work since it would cost us a lot of data points, therefore, we had to design an algorithm for doing it as accurately as possible. The code for the final algorithm will be on Github in branch

The second part of cleaning was to work on categorical columns. The issue was, that for most categorical columns in the given format should've been one of two variants they given, they actually had many more, where some of the different variants, caused a different issue. So, we started by looking through each column, starting with `zip_code`, most of them were fine, however, some were under different formatting, so we fixed that. Then we worked on `at_home` and `extra_large_column`, they also had a lot of different values so we just mapped the values into a domain of 3 values. However, two columns `level_of_study` and `full_time` were in a worse of state than others. While inspecting the data we realized that some rows had values such as `3 UG; 5 G` or `all FT` for some rows. Which actually meant 8 students residing on the same addres, 5 of them being graduates, and the other 3 undergraduates, all studying full-time. This issue appeared in a few thousand entries, which we didn't want to lose so we fixed that.

Right now we are still looking for the connections between datasets, but we did find a few already, and started cleaning the ones we found.

However, skimming over a data we hope to try to build neighborhood ranking and house ranking to try to track changes in off-campus student housing. Possibly identifying any correlations and visualizing them to analyze.

Notebook for merging: `fa24-team-b/data_prep_phase/data_prep.ipynb`

Script for address_parser: `fa24-team-b/data_prep_phase/address_parser.py`

Notebook for data cleaning: `fa24-team-b/data_prep_phase/merge_and_clean.ipynb`

Workload

- Raul

    1. Inspecting the initial dataset and merging a subgroup of the full dataset.
    2. Writing an algorithm for address parsing and applying it on addresses that were not separated
    3. Cleaning the rest of the merged dataset.

- Zainab

    1. Inspecting the data, cleaning the data, and developing code to merge all the clean data of all schools.
    2. Inspecting “Property Assessment” datasets and looking for possible ways to connect it to the main student addresses dataset.
    3. Started cleaning 2024 Property Assessment dataset.

- Christine

    1. Inspecting the dataset and merging a part of the data into a group.
    2. Inspecting other datasets for possible connections to the main one.
    3. Started the cleaning of the building and property violations dataset.

- Gabby

    1. Inspecting the dataset, cleaning the data and merging each year of the dataset into a dataset, grouped by school.

2. Below is an example of visualization of student addresses for one of the Boston schools (Wentworth Institute of Technology)