Deliverable 2 reprot

Collect and Pre-process data, preliminar analysis

We finally got the clients' data, but the data was very massy. We first tried to clean the data (i.e. fix some disposition and etc), and further clean is still needed to be done. Then based on the cleaned data, we analyzed the distribution of the client population. In addition, we geocode the clients' address into coordinates, for future use of analyzing census tract.

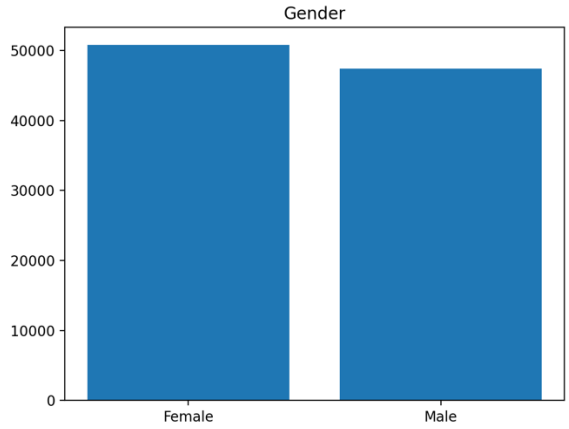
Current work

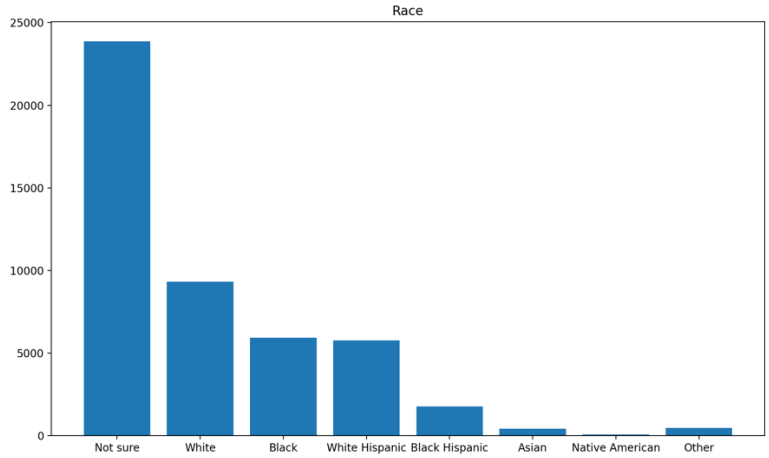
1. Try to clean the data, but more clean needed to be done.
2. Draw frequency distribution of races, gender, location, etc.
3. Geocoding the clients' address into coordinates, and relating each coordinate to a census tract.
4. Plot the school, counthouses, and YAD offices on the map.

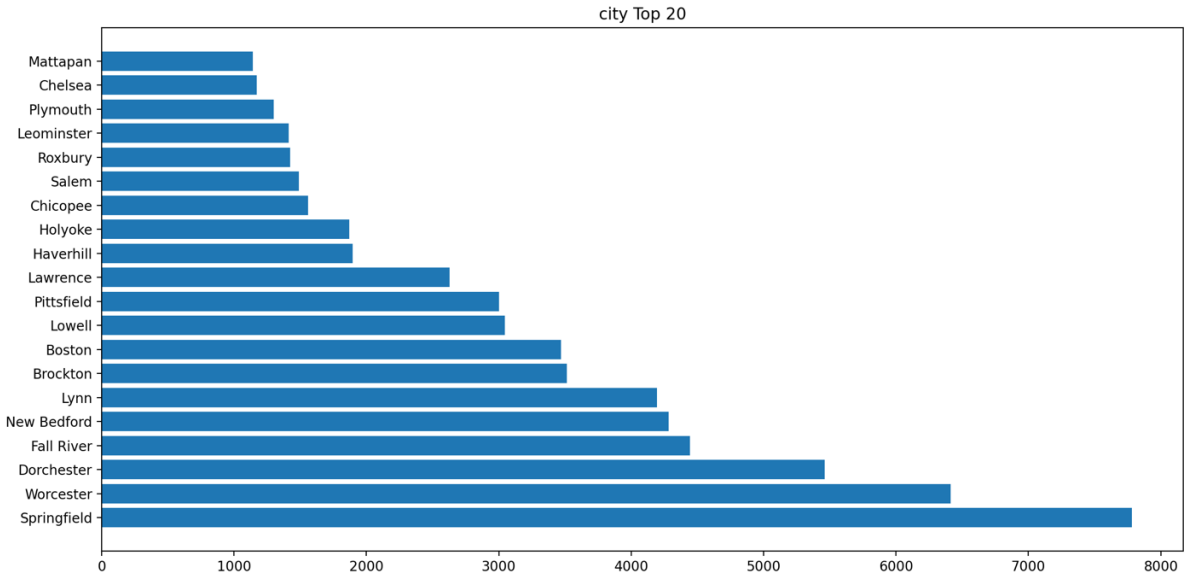
Key questions

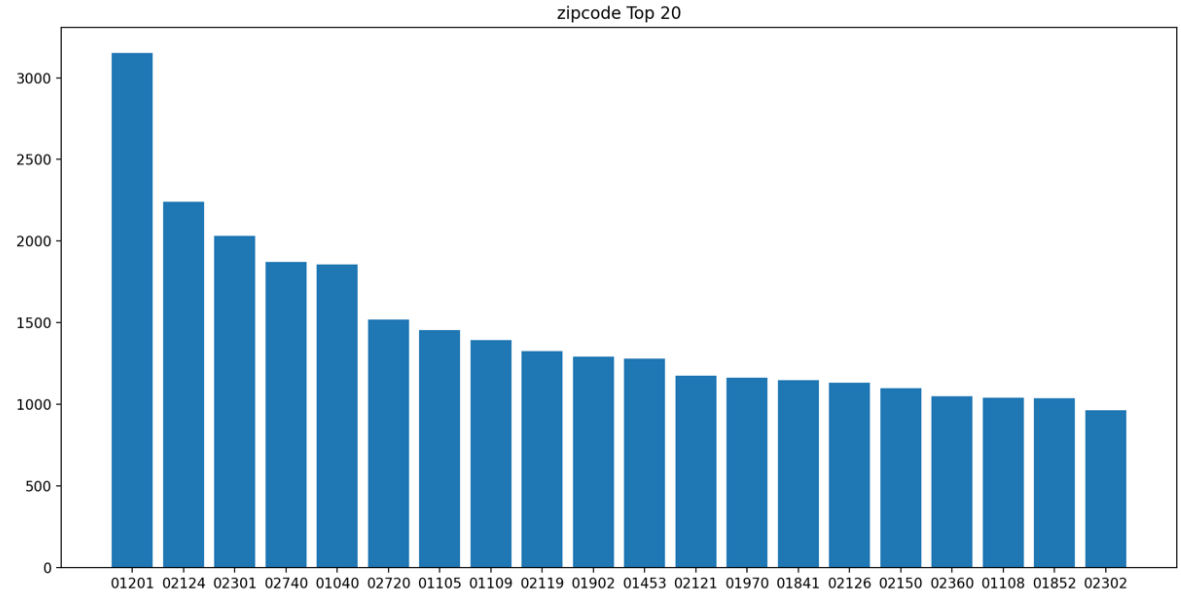
1. For clients in the dataset, what's the proportion of the races, genders, locations, etc?

   We plotted the following graphs for preliminar analysis:









2. How good is the dataset? Is it well-formatted and easy to do analysis on?

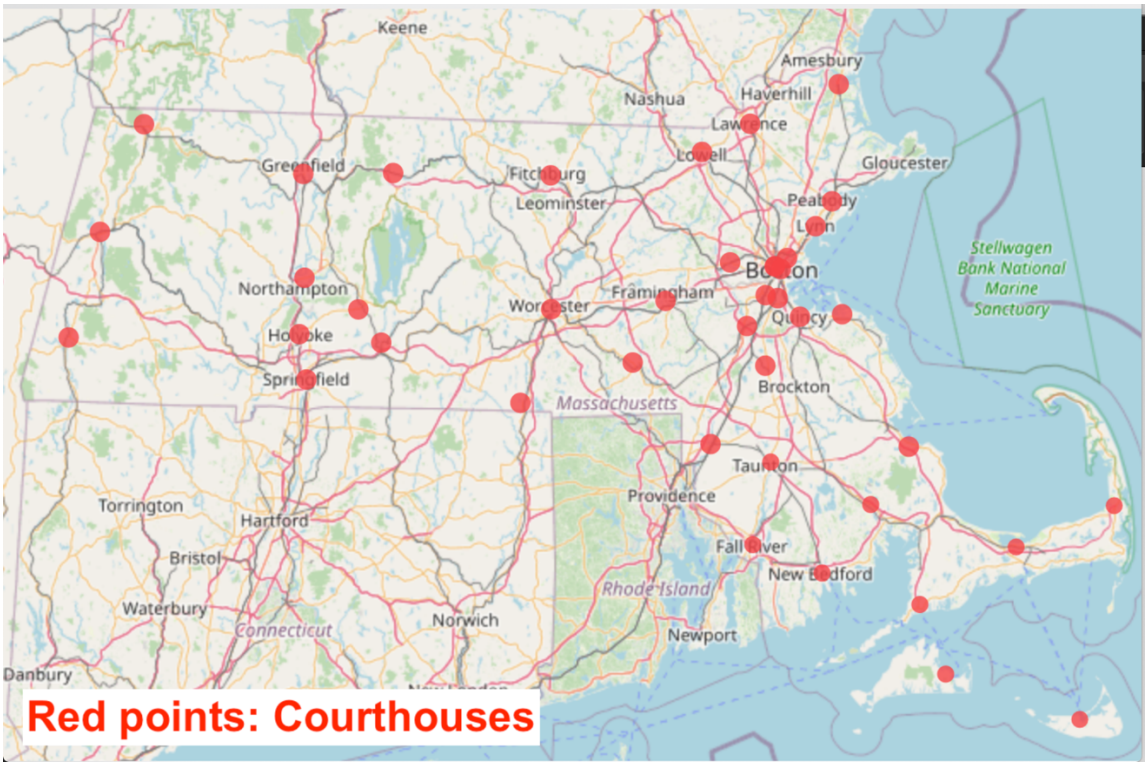
   When we received the data, we found that the data was messy and there were many misalignment problems. After observation, we believe that the data is messy because when the client input information into database it contains commas, and then when database information transit into a CSV file, the comma would be separated, which caused a lot of data misalignment.

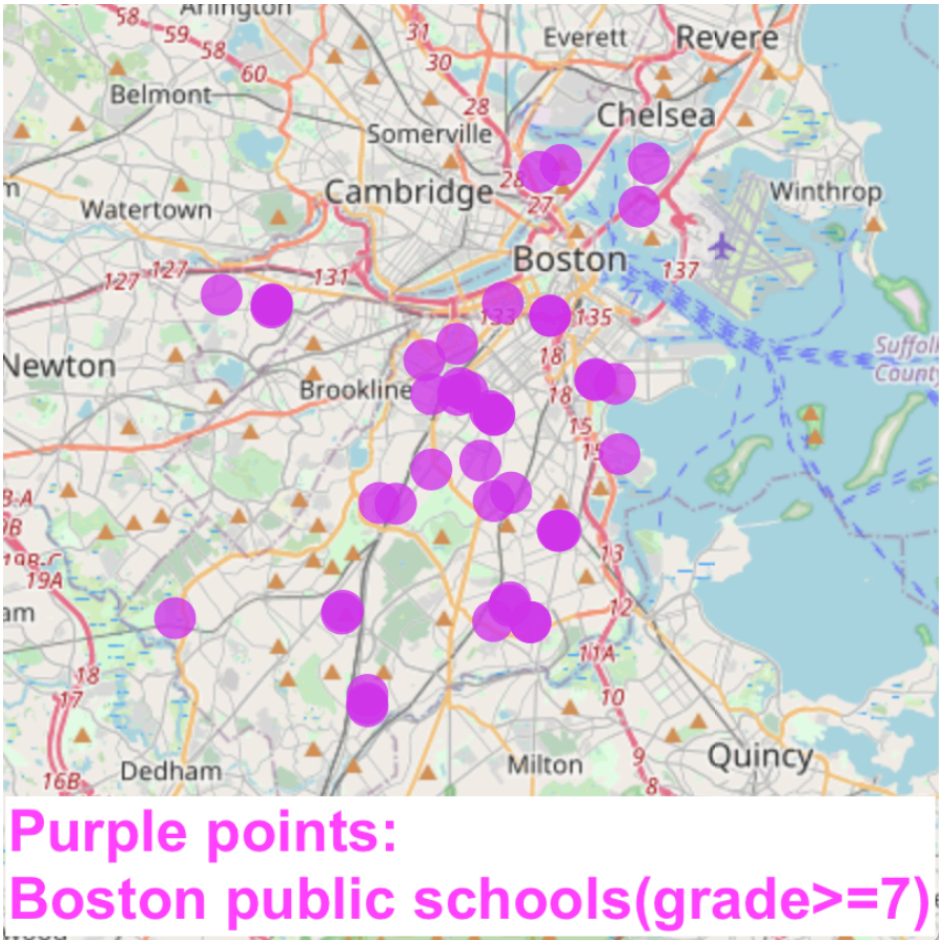
As a whole, we now have a `csv` file with `180823 rows × 34 columns`. And we are waiting for our client to provide further more data.

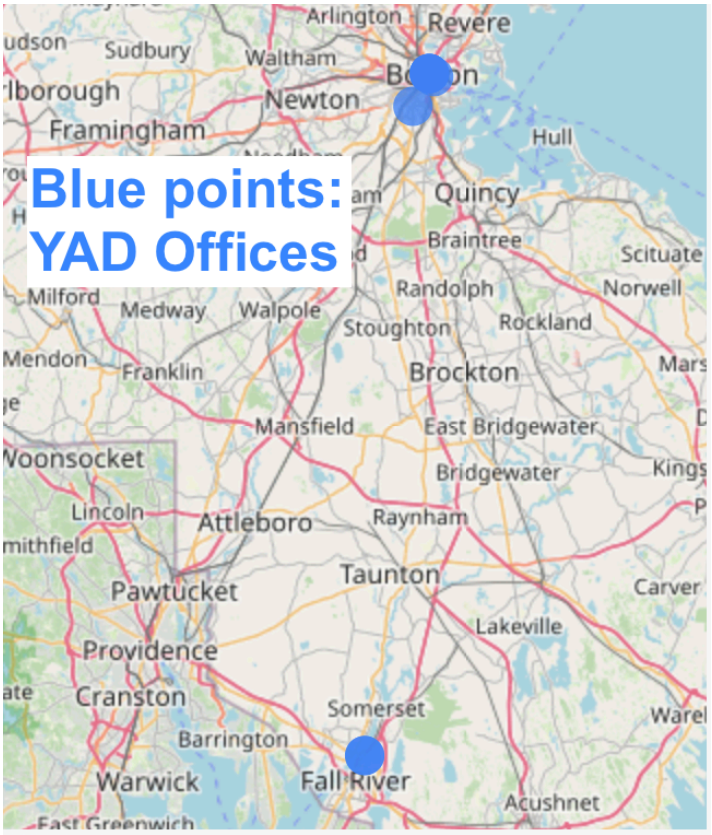
Column Name(Property)	# of NaNs	NaN Rate
person_id	/	/
dob	83265	46.0%
intake_date	14689	8.1%
address_line 1&2&3	48148	26.6%
city	54954	30.4%
state	51	0.0%
zip	83318	46.1%
Gender	82574	45.7%
Race	133157	73.6%
Ethnicity	135704	75.0%
Primary lang ID	156345	86.5%
Primary Language	156345	86.5%
Spoken Lang IDs	169262	93.6%
CPCS Office	107736	59.6%
place_of_birth	169199	93.6%
rec_entered_date	1268	0.7%
si_dcf	174214	96.3%
si_dmh	174214	96.3%
si_dds	174214	96.3%
si_dph	174214	96.3%
si_dta	174214	96.3%
si_cbhi	174214	96.3%
si_mrc	174214	96.3%
si_cp_atto	174214	96.3%

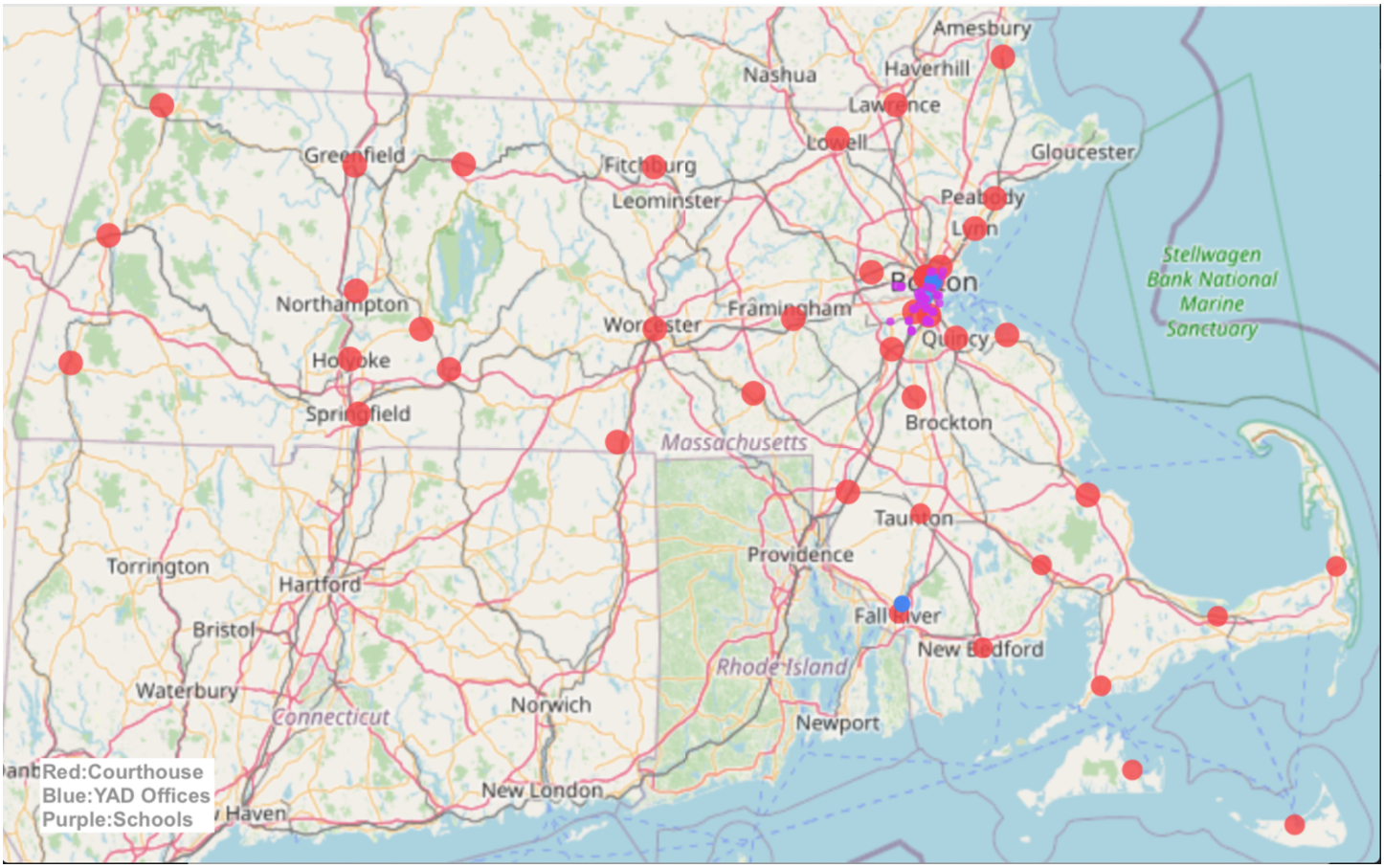
si_chins_atto	174214	96.3%
si_gal	174214	96.3%
si_chins_probation	174214	96.3%
si_crim_probation	174214	96.3%
si_crim_atto	174216	96.3%
si_others	174878	96.7%
Defense Attorney	179460	99.2%
IsHispanic	123146	68.1%

3. How are schools, YAD offices, and courthouses distributed in Boston areas?






Blue points:
YAD Offices
Arlington
Revere
Sudbury
Waltham
Newton
Framingham
Hull
Quincy
Braintree
Scituate
Randolph
Norwell
Milford
Medway
Walpole
Stoughton
Rockland
Mendon
Franklin
Brockton
Mansfield
East Bridgewater
Bridgewater
Lincoln
Attleboro
Raynham
Taunton
Carver
Pawtucket
Lakeville
Providence
Cranston
Somerset
Barrington
Warwick
Fall River
Acushnet


Amesbury
Nashua
Haverhill
Lawrence
Lowell
Gloucester
Greenfield
Fitchburg
Leominster
Peabody
Lynn
Stellwagen
Bank National
Marine
Sanctuary
Northampton
Framingham
Worcester
Quincy
Holyoke
Springfield
Brockton
Massachusetts
Taunton
Providence
Torrington
Hartford
Fall River
Bristol
New Bedford
Rhode Island
Waterbury
Connecticut
Norwich
Newport
New London
Red:Courthouse
Blue:YAD Offices
Purple:Schools