MFLP 90803 – Data Cleaning Read Me

Chi-Shiun Tsai and Colton Lapp

Project Title:

Understanding Urban Depopulation: An analysis of shrinking cities in the US with machine learning approaches

Intro:

This readme explains the:

- Data sources
- Data cleaning steps
- Procedure for recreating our final datasets

Data Sources:

City Data - American Community Survey (ACS): 5 year data

Source: Census API via the Census python package

- **Year**: 2019, 2020, 2021
- **Variables** ~ 40 (Demographic data, income data, home value data, etc)
-**Geography**: "City" level (*places* in the Census lingo)
- Python Package URL

COVID Data

Source: NYTimes Covid Tracking Github Page

- **Year**: 2020-2023
- **Variables** ~ Cases and Deaths
- **Geography**: County level
-URL

Crime Data

Source: FBI Crime Data Explorer

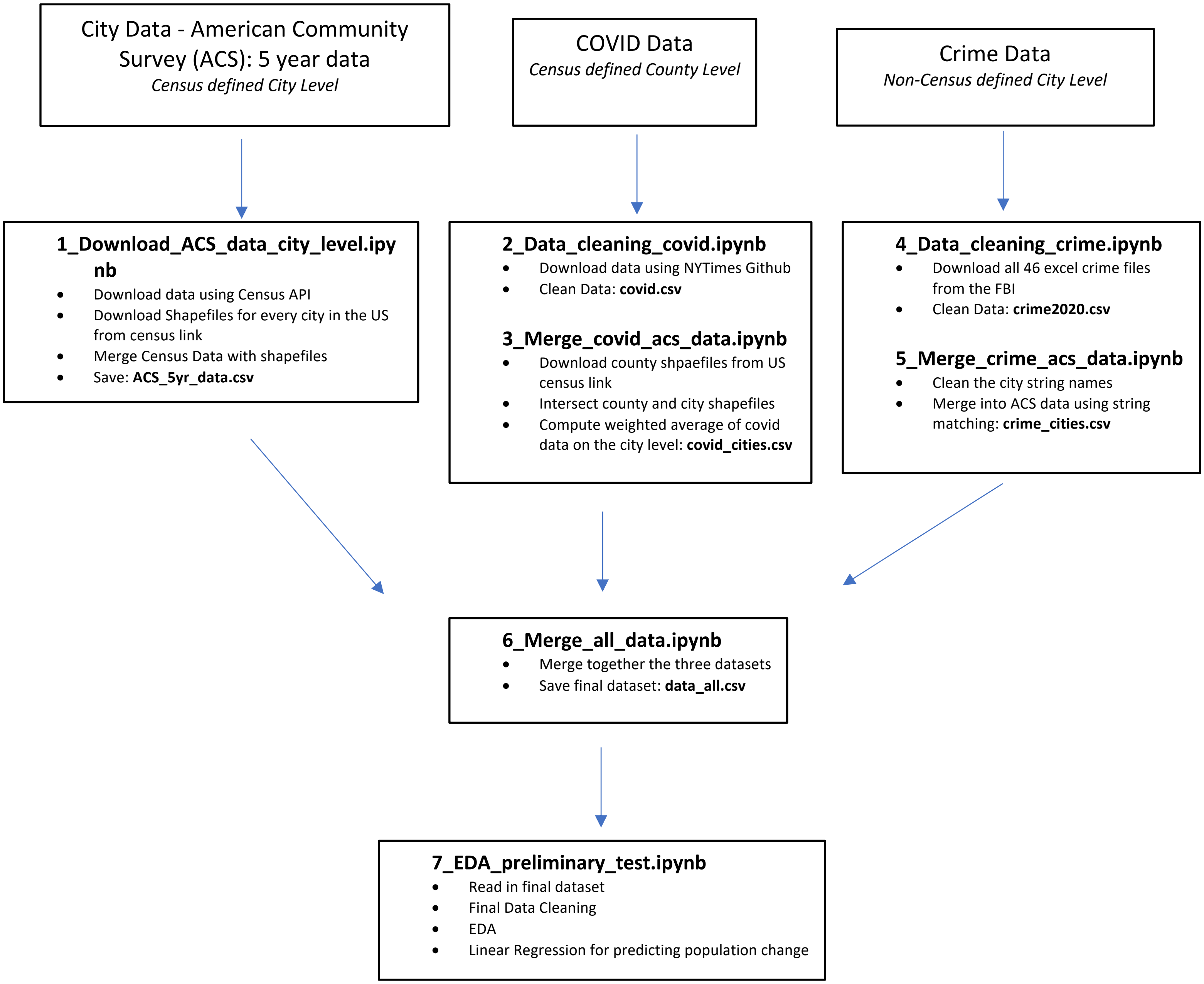
- **Year**: 2020
- **Variables** ~ Incidents
- **Geography**: City level (different than Census defined cities)
- URL

Data Cleaning:

The hardest part of our data cleaning was getting all the data on the same geographic level. We wanted our data to be on the city level because that is the most intuitive for people to understand (people think of "New York" as a single place, not as a combination of 21 counties). The issue with turning all our data into the city level is that there is not a well standardized method of classifying city boundaries. The official city boundaries often times are much smaller than the Metropolitan Statistical Area that surrounds it. Some cities also span multiple states (i.e. Kansas City).

The way we solved this was to use GIS processing. We merged the ACS census data with census shapefiles. This gave us a geographic boundary of every city. We then projected COVID data on the county level to the city level by intersecting the county and city geographies. For example, if one city resided in two counties, we made the COVID data of that city be equal to the average of the two counties it resides in.

The crime data from the FBI was already on the city level, but it did not use the same city definitions as the census. Because the city classification systems differed, we tried our best to merge the data using string matching of city names. Despite our best efforts, there were still many cities that we could not match data to. The FBI dataset also did not include some important major cities, such as NY City and Los Angeles.



City Data - American Community Survey (ACS): 5 year data
Census defined City Level

COVID Data
Census defined County Level

Crime Data
Non-Census defined City Level

1_Download_ACS_data_city_level.ipynb

- Download data using Census API
- Download Shapefiles for every city in the US from census link
- Merge Census Data with shapefiles
- Save: **ACS_5yr_data.csv**

2_Data_cleaning_covid.ipynb

- Download data using NYTimes Github
- Clean Data: **covid.csv**

3_Merge_covid_acs_data.ipynb

- Download county shpaefiles from US census link
- Intersect county and city shapefiles
- Compute weighted average of covid data on the city level: **covid_cities.csv**

4_Data_cleaning_crime.ipynb

- Download all 46 excel crime files from the FBI
- Clean Data: **crime2020.csv**

5_Merge_crime_acs_data.ipynb

- Clean the city string names
- Merge into ACS data using string matching: **crime_cities.csv**

6_Merge_all_data.ipynb

- Merge together the three datasets
- Save final dataset: **data_all.csv**

7_EDA_preliminary_test.ipynb

- Read in final dataset
- Final Data Cleaning
- EDA
- Linear Regression for predicting population change

Voting Data

Originally, we hoped to combine voting data into our dataset. We downloaded voting data at the precinct level and cleaned it. We were unable to join the voting data to the city level, however, because the geographies of precincts are too complicated and divergent to systematically join to the city level. You can read more about this issue here: https://redistrictingdatahub.org/data/about-our-data/election-results-and-precinct-boundaries/

There is a jupyter notebook file we left in the repo called data_cleaning_covid.ipynb which you can look at if you want to see what data cleaning we completed. We may look into this issue later on if we have time, but for now, we believe we have sufficient variables.

Recreating our final dataset:

To recreate our final dataset and then do preliminary EDA and modeling, run these scripts in order:

- **1_Download_ACS_data_city_level.ipynb**
- **2_Data_cleaning_covid.ipynb**
- **3_Merge_covid_acs_data.ipynb**
- **4_Data_cleaning_crime.ipynb**
- **5_Merge_crime_acs_data.ipynb**
- **6_Merge_all_data.ipynb**
- **7_EDA_preliminary_test.ipynb**

A description of what each script does is available in the markdown of each file as well as above in the procedure flow chart.

Note on raw data storage/LFS in Github:

The shapefiles for cities and counties take up a lot of space (shapefile for cities is around 200MB). As a result, we cannot store them in Github. We have tried to modify our scripts to not save the shapefiles but instead download them directly when needed in every script. The raw covid data files are also very large (all combined are around 200-400MB). Because of this, we tried to store these files in google drive. We were having issues reading google drive files into our jupyter notebooks, however, especially with the shapefiles. As a result, we had to read these files in locally. If you still have this issue, you can download these files from our google drive folder here and read them in locally as well:

https://drive.google.com/drive/folders/1Ho3Ufsi9GW8Md1rqI_P0sHjxDC-0ADFG?usp=share_link