#Iterations

Iteration 1:

Data was cleaned.

Income data was cleaned and concatenated into one database. We only kept year, county and income columns.

Race data was cleaned and concatenated in one database, and different races we kept were added together from male/female to total.

Age/Sex was cleaned and concatenated in one database, age groups, age median, male/female ratio, counties and years will work.

Rural vs Urban counties was kept the same.

Needs: Find a way to transpose Zillow data into rows.

Iteration 2:

Originally we got data for the years 2017-2021 after the first round of ML we removed variables.

All databases were made sure to have either a FIPs, countyname,Statename, or countyname,StateAbrivation so that FIPS database can be used to join all databases. After this we have not yet removed the unneeded columns.

ERD:

![ERD](https://user-images.githubusercontent.com/109693301/207495541-33df45d6-d182-4eb1-b512-86207fc6ce04.png)

All tables are connected by fips, fips by year, and fips by year and month. This ERD only contains columns we are going to keep.

For the housing cost Zillow data we used pd.stack() which allowed us to have all cost data to be in rows and indexed by their fips. This allows us to convert time series data into individual instances of data for each county. We went in and manually changed the names of rows because the issue with stack was that we were not able to manipulate the data after. Finally, we broke the dates into months and years to make the fips by year and fips by year and month above. This table is currently under Zillow\_int2.csv

All examples of these iterations are currently in the Circle branch.

Needs: Merge data using SQL, and finish cleaning data.

Iteration 3:

We have cleaned all of the data individually removing all Nan when merging. We have merged the data into 2 dataframes, housing/covid data and demographic (population, race/eth, income, and Rural/urban areas).

We uploaded these 2 dataframes to Mongo as collections using PyMongo.

Needs: Merge these two frames in Mongo.

Iteration 4:

Pivot I merged the data in SQL not Mongo as it was not working. The query is schema.sql. The data was then downloaded and uploaded into Mongodb (currently in the github as mergedata.csv in clean data).

Mongodb has been found to be shared over the cloud via Atlas. This is currently being tested.

Needs: Mongodb needs to be shared with others.

Iteration 5:

As "fips" is needed for the tablau we redid the merge in SQL to include this in the database. The data was then downloaded and uploaded into Mongodb atlas. We then shared this Atlas with everyone in the group.

After all of this was finished we realized that this data set did not have COVID cases for New York City. This was then removed leaving the final amount of rows being 95131.