Covid-19 Dataset Overview
Project Group 1
Phase 1

Derek Fox, David Mack, Trevor Church, Sarah Barber

Merging Super Covid

- The super_covid_data.csv file contains three datasets
  - covid_deaths_usafacts.csv
  - covid_cases_usafacts.csv
  - covid_county_population_usafacts.csv
- Each of these datasets contains similar information
  - countyFIPS – A federal identifier for counties
  - County Name – The name of the county
  - State Name – The name of the state
  - StateFIPS – The first two digits in the FIPS

Name	Definition	Data Type	Example Values	Required?
countyFIPS	County ID Number	Integer	1000, 1001, ...	yes
County Name	Name of County	String	Bibb County, Coffee County, ...	yes
State	State Abbreviation for County	String	AL, NC, ...	yes
State FIPS	State ID Number	Integer	1, 2, ...	yes
Cases – by Date	Number of Cases on given Date	Integer	0, 100, ...	yes
Deaths – by Date	Number of Deaths on given Date	Integer	0, 100, ...	yes
Population	County Population	Integer	1000, 5000, ...	yes

Merging Super Covid

Given how the information in the spreadsheets is organized merging is quite simple.

- Merge the Cases and Deaths dataframes to joint ‘cases_deaths’ dataframe
  - Add suffixes to date columns to specify case or death information
- Add on county population data to create ‘super_covid’ dataframe
- Finally, export to csv

```
cases_deaths = pd.merge(cases, deaths, on=['countyFIPS', 'County Name', 'State', 'StateFIPS'],
suffixes=('_cases', '_deaths'))
```

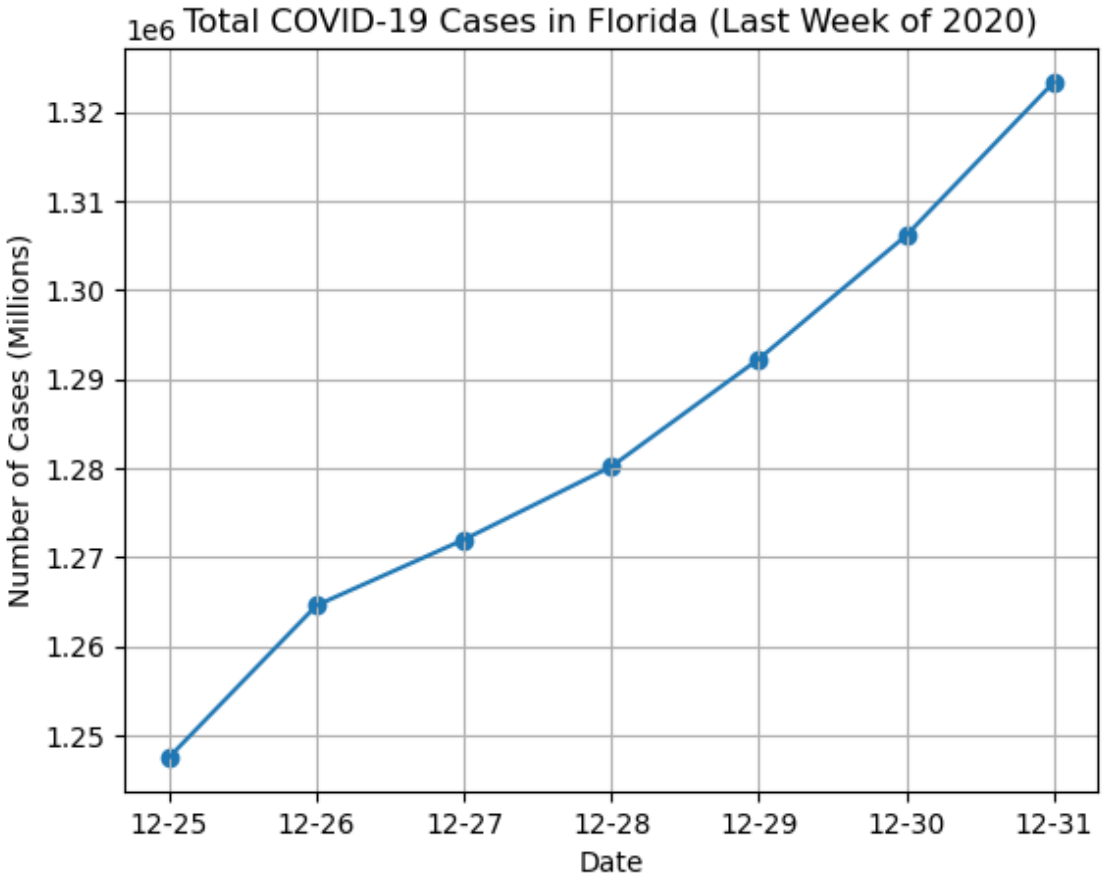
```
super_covid = pd.merge(cases_deaths, county_pop, on=['countyFIPS', 'County Name', 'State'])
```

```
super_covid.to_csv('./data/super_covid_data.csv', index=False)
```

COVID-19 Florida Statistics - Derek Fox

- Data pipeline:

- Plot millions of cases vs. date
- Found that cases in last week of 2020 were increasing



Presidential Election Dataset – Derek Fox

Data Dictionary

Name	Definition	Data Type	Example Values	Required?
state	State Name	String	Florida, Alabama, ...	yes
county	County Name	String	Orange County, Washington County, ...	yes
candidate	Candidate Name	String	Joe Biden, Donald Trump, ...	yes
party	Abbreviation of Party Name of Candidate	String	DEM, REP, ...	yes
total_votes	Number of Votes for Candidate	Integer	1000, 5000, ...	yes
won	Did Candidate Win?	Boolean	True, False	yes

Presidential Election Dataset – Derek Fox

- Merge on “County” and “State”
  - Problem: “State” column in covid data is abbreviation, not full name
  - Solution: Map from state name to abbreviation
  - Now ready to merge!

Note: this variable is just a python dictionary with mappings from state names to their abbreviations e.g. “Florida”: “FL”

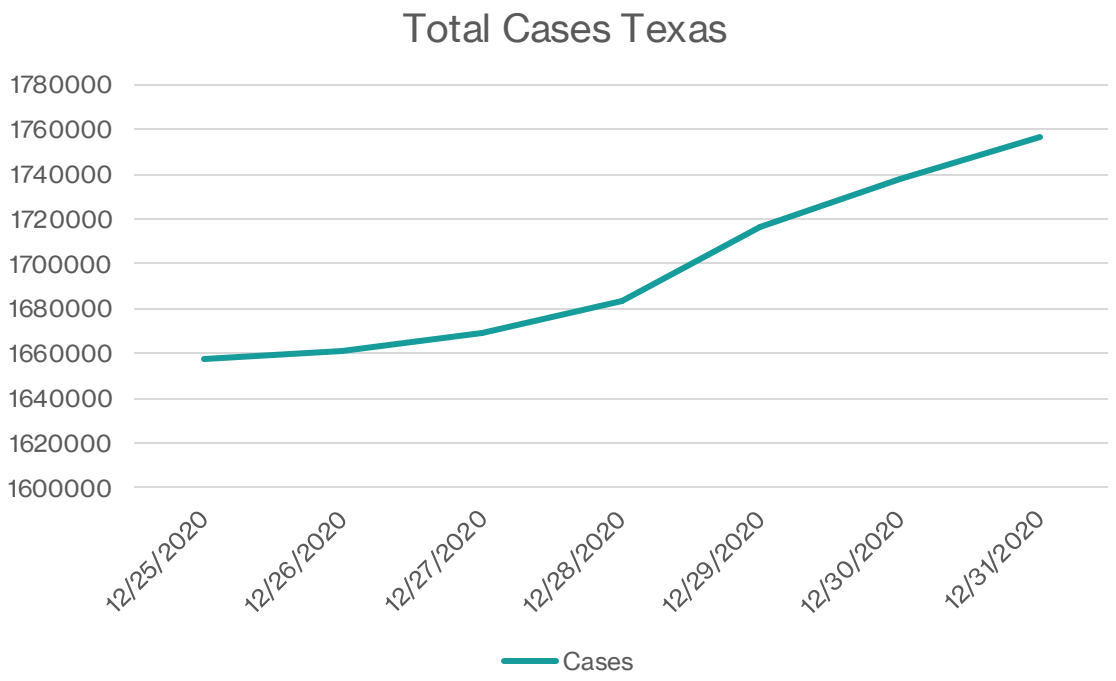
```
# Add state abbreviation column
pres_by_county['state_abbr'] = pres_by_county['state'].map(state_name_to_abbr)
```

```
# Merge the two dataframes on the state abbreviation and county columns
merged = pd.merge(covid_2020_data, pres_by_county, left_on=['State', 'County Name'], right_on=['state_abbr', 'county'])
```

- Hypotheses for next stage:
  - Higher mortality rates (deaths per *x* cases) in counties that voted for Donald Trump
  - Higher voter turnout correlated with more deaths

COVID-19 Texas Statistics – David Mack

- Locate the last week of COVID-19 case statistics.
- Cull super_covid and only include that data.
- Cull this new dataset so it only includes counties in Texas
- Sum cases each county over the week to create your total.



Employment Dataset - David Mack

- The Bureau of Labor Statistics publishes employment data.
- This data is classified by FIPS among other things. This will be my merge column.
- FIPS in the BLS data is a 5-digit code in the form 0XXXX or XXXXX
- The BLS FIPS must be typecast before the merge as a result.

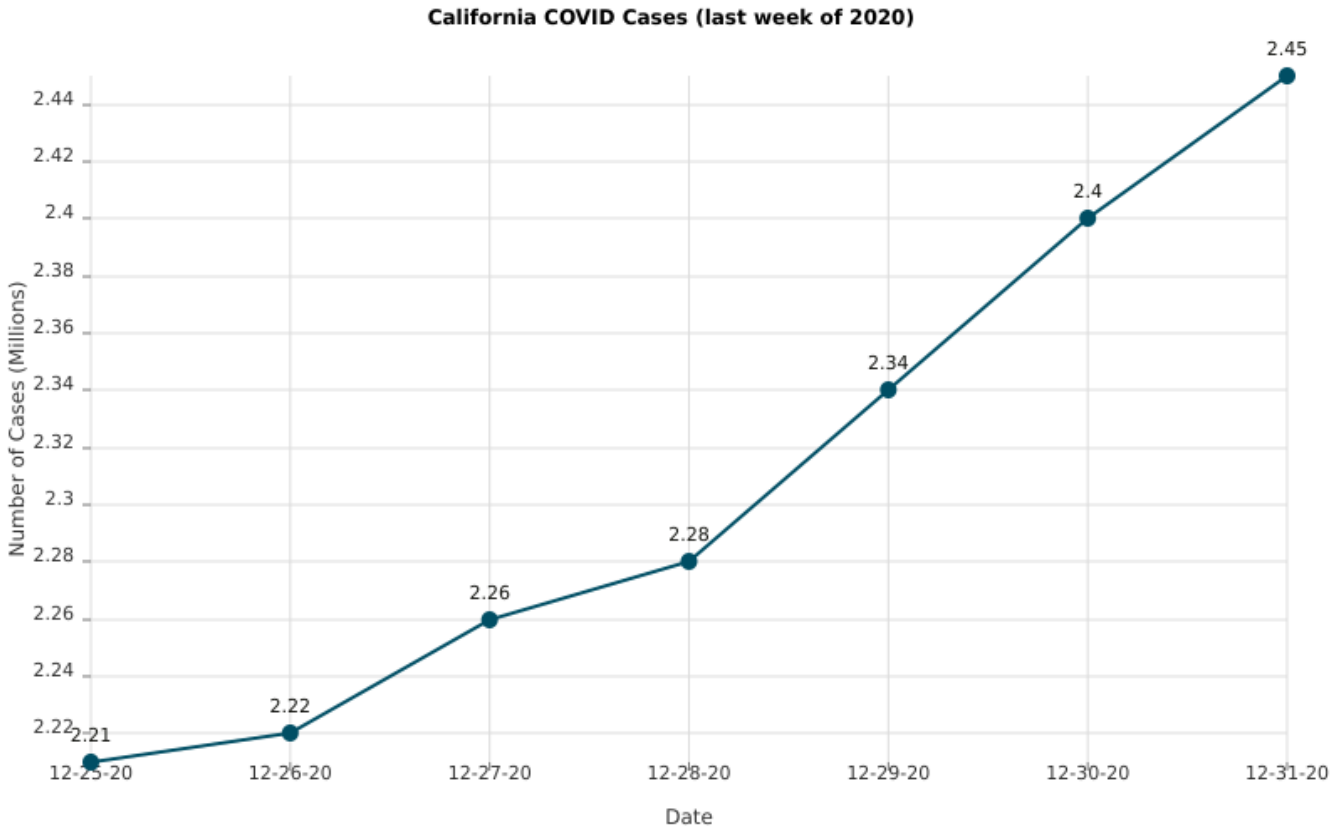
```
bls_cull = bls_data[bls_data['Area Type'] == 'County']

# Change the Super Covid and BLS datasets to use integers for FIPS for easy handlin
bls_cull['Area\nCode'] = bls_cull['Area\nCode'].astype('int64')

# Rename the column so that I can merge
bls_cull = bls_cull.rename(columns={'Area\nCode':'countyFIPS'})

# Merge The Dataframes
covid_employment = pd.merge(super_covid, bls_cull, on=['countyFIPS'], how='left')

covid_employment
```

Employment Data - Hypothesis

- There is a correlation between average income and cases.
  - Locations with a higher average income tend to be within cities and I believe that cities will have more infections due to proximity.
- There is a correlation between average income and deaths.
  - When you control for total cases. I believe that the death rates will be higher in locations with lower income due to lower access/utilization of medical care.

Covid-19 California Statistics – Trevor Church

- Found that cases in the last week of 2020 were increasing
- The super covid sheet was filtered by year, state, and cases, and then inspecting the final 7 entries.
  - I added a 'totals' row once the data had been filtered to 7 columns, and had it total the cases from all counties each day
  - From the 25th to 31st of December in 2020, Covid cases in California rose from 2,213,261 to 2,455,296



Educational Attainment Dataset – Trevor Church

- Merge on "County Name"
  - The educational attainment set has a column "NAME" with 'county name, state name', listed alphabetically by county
  - The super covid set has a column "County Name" with just county names listed alphabetically
  - If we rename the column and drop state names, it will be easy to merge

```
update = update.rename(columns={'NAME': 'County Name',
```

'update' is the updated dataframe

```
update['County Name'] = update['County Name'].str.split(',').str[0]
```

Removing all state names and commas

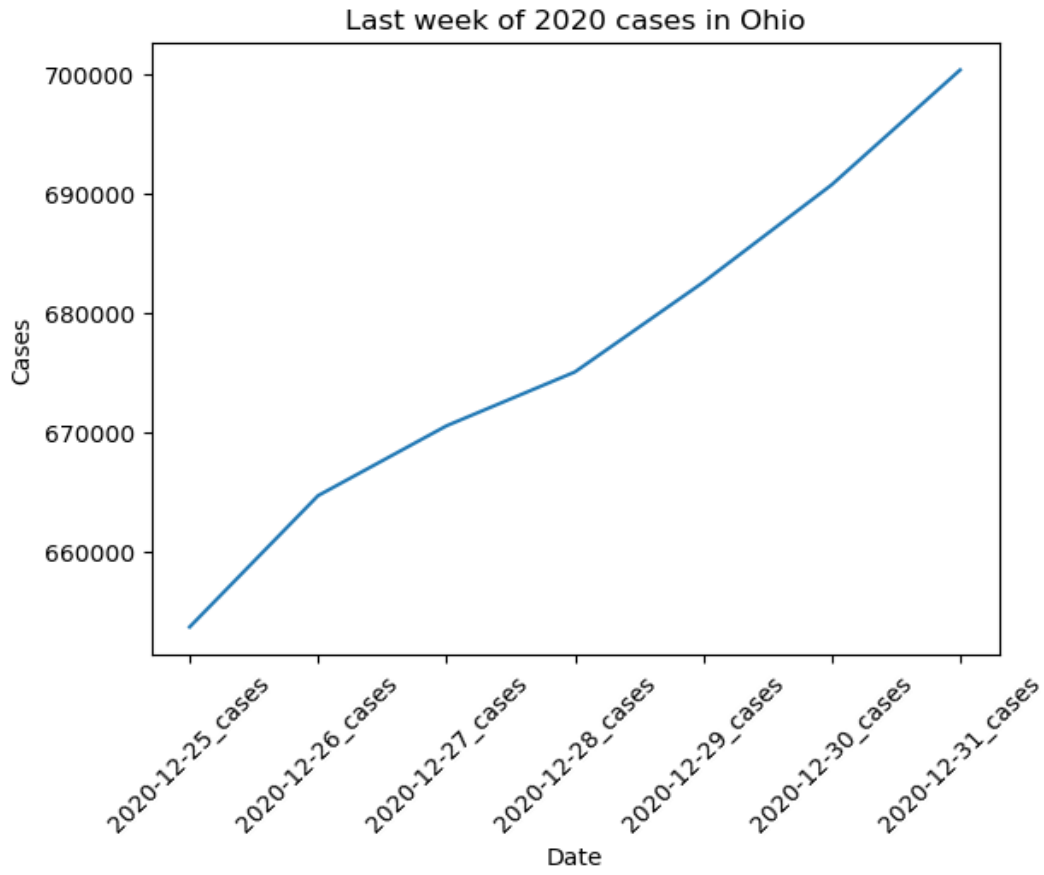
```
merged = pd.merge(update, covid2020, on='County Name')
```

Merged with 2020 covid data, as the census is from 2020

- Hypothesis for next stage: Counties with lower rates of educational attainment will have higher Covid cases

Covid-19 Ohio Statistics – Sarah Barber

- Results indicated that cases increased over the course of a week by going off the number of cases in each county.
  - Obtained by filtering down to only 2020.
  - Then filtered to only data for Ohio.
  - Finally, filtered to the last week of 2020.
  - Results were displayed as a graph.



United States Census Dataset

- Consists of data on the estimated total populations of each county in every state in 2020. This data is sorted by age ranges and gender as both a total number and a percentage.

County Name	State	Estimate Total Population	Estimated Pop. <5	Estimated Male Total Population	Estimated Female Total Population	Etc.
Randolph County	North Carolina	143460	8086	70830	72630	...

Merging Census Dataset

- Data is merged on the "County" and "State" columns
  - Issue: Census dataset is not predisposed to making county and state columns.
  - Solution: Manually rework the dataset to include these two columns and eliminate redundancy.

```
df = pd.read_csv('Census_Enrichment.csv')
covid = pd.read_csv('super_covid_data.csv')

merged = pd.merge(df, covid, on=['County Name', 'State'], how='left')

merged.head()
print(merged)

merged.to_csv('covidandreducedcencus.csv')  #This converts the merged dataframe into a new csv
```

  - *Hypothesis: Counties/States with a higher population of older people may have more recorded deaths and states with a higher overall population may have a higher death rate.*