**Effects of Wearing Masks on COVID-19 Cases (by county)**

Team Members:

*Aleena Ghumman*

*Emerson Williams-Molett*

*Paul Bernhardt*

*Matt Killeen*

**Requirements**:

* Describe the core message or hypothesis for your project.
* Describe the questions you and your group found interesting, and what motivated you to answer them
* Summarize where and how you found the data you used to answer these questions
* Describe the data exploration and cleanup process (accompanied by your Jupyter Notebook)
* Describe the analysis process (accompanied by your Jupyter Notebook)
* Summarize your conclusions. This should include a numerical summary (i.e., what data did your analysis yield), as well as visualizations of that summary (plots of the final analysis data)
* Discuss the implications of your findings. This is where you get to have an open-ended discussion about what your findings "mean".
* Tell a good story! Storytelling through data analysis is no different than in literature. Find your narrative and use your analysis and visualization skills to highlight conflict and resolution in your data.

**Presentation Requirements:**

You are free to structure your presentations to your liking, but students tend to have success with the following format.

* Title Slide
  + Include the name of the Project and Group Members
* Motivation & Summary Slide
  + Define the core message or hypothesis of your project.
  + Describe the questions you asked, and *why* you asked them
  + Describe whether you were able to answer these questions to your satisfaction, and briefly summarize your findings
* Questions & Data
  + Elaborate on the questions you asked, describing what kinds of data you needed to answer them, and where you found it
* Data Cleanup & Exploration
  + Describe the exploration and cleanup process
  + Discuss insights you had while exploring the data that you didn't anticipate
  + Discuss any problems that arose after exploring the data, and how you resolved them
  + Present and discuss interesting figures developed during exploration, ideally with the help of Jupyter Notebook
* Data Analysis
  + Discuss the steps you took to analyze the data and answer each question you asked in your proposal
  + Present and discuss interesting figures developed during analysis, ideally with the help of Jupyter Notebook
* Discussion
  + Discuss your findings. Did you find what you expected to find? If not, why not? What inferences or general conclusions can you draw from your analysis?
* Post Mortem
  + Discuss any difficulties that arose, and how you dealt with them
  + Discuss any additional questions that came up, but which you didn't have time to answer: What would you research next, if you had two more weeks?
* Questions
  + Open-floor Q&A with the audience

**Technical Requirements:**

* Use Pandas to clean and format your dataset(s)
* Create a Jupyter Notebook describing the **data exploration and cleanup** process
* Create a Jupyter Notebook illustrating the **final data analysis**
* Use Matplotlib to create a total of 6-8 visualizations of your data (ideally, at least 2 per "question" you ask of your data)
* Optionally, use at least one API, if you can find an API with data pertinent to your primary research questions
* Create a write-up summarizing your major findings. This should include a heading for each "question" you asked of your data, and under each heading, a short description of what you found and any relevant plots.
* Your project must include at least one testable hypothesis or at least one linear regression.

**Core Message**: We aim to determine the correlation between self-reported mask-wearing behavior per U.S. county and COVID-19 case rate and death rate per county as of July 14, 2020.

1. **The questions we found interesting:**
   1. Is there a correlation between COVID *cases* per 100,000 pop. and:
      1. lower mask-wearing score?
      2. higher population density?
   2. Is there a correlation between mask-wearing score and population density?
   3. Is there a correlation between COVID *deaths* per 100,000 pop. and:
      1. lower mask-wearing score?
      2. higher population density?
2. **Findings:**
   1. Is there a correlation between COVID *cases* per 100,000 pop. and:
      1. lower mask-wearing score? *There is a correlation, but it’s in the opposite direction we thought! There are more cases in counties with higher mask-wearing scores.*
      2. higher population density? *There are more cases in counties with higher population densities.*
   2. Is there a correlation between mask-wearing score and population density? *Counties with higher population densities self-reported better mask-wearing scores.*
   3. Is there a correlation between COVID *deaths* per 100,000 pop. and:
      1. lower mask-wearing score? *There is a correlation, but again it’s in the opposite direction we thought. There are more deaths in counties with higher mask-wearing scores.*
      2. higher population density? T*here are more cases in counties with higher population densities.*

Group discussion point

The last three digits of a FIPS code represent county. The first two digits (or one digit) represent state. Different data sets handle state by using (or not using) a leading zero for state codes 01 Alabama, 02 Alaska, 04 Arizona, 05 Arkansas, 06 California, 08 Colorado, and 09 Connecticut. The first county of the first state on the FIPS list, Autauga County, Alabama, is represented in various datasets as either 01001 or 1001.

Census Bureau, land mass, and *NYT* mask-use all round the leading zero, while *NYT* case/death data use the leading zero. How do we want to make FIPS consistent across all intermediary dataframes so their select contents can be merged happily into the GBDf?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *dataframe* | *name* | *Autauga* | *type* | *records* |
| **0** | **GBDf** | **FIPS** | **01001** | **int** |  |
| 1 | Census data | FIPS | 1001 | int | 3142 |
| 2 | caseData | fips | 1001.0 | float | 3136-24=3112 |
| 3 | landMassData | Areaname | 1001 | int | 3220 |
| 4 | maskWearingData | COUNTYFP | 1001 | int | 3142 |
| 5 | county centers | fips | 1001 | int | 3147 |

The simplest way to make FIPS consistent across dataframes is to make it an integer in all dataframes. This should only require one change, in the *NYT* case data.

As of July 14, 2020, 23 rows in the cases and deaths dataset had no county associated with it. These represented a total of 14,402 cases and 483 deaths, out of 3,433,765 cases and 136,173 deaths that had happened cumulatively up until July 14, 2020, or 0.42% of cases and 0.35% pct of deaths.

We’re doing an inner merge, with a little bit of extra work on New York City. The dataset leaves zeros in the cases and deaths column for the five counties that make up New York City, but provides aggregate cases and deaths figures for the whole city. We should create a placeholder FIPS code, “99999,” for New York City, that allows population, mask-wearing scores, and land mass for the five counties to be added together:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *borough* | *county* | *FIPS* | *pop.* | *Land mass (sq.mi)* | *Pop. density* | *Mask-wearing score* | *Geogr. center (lat.,long.)* |
| Bronx | Bronx County | 36005 | 1,418,207 | 42.05 | 33,727 | 8.80 |  |
| Brooklyn | Kings County | 36047 | 2,559,903 | 69.82 | 36,664 | 8.75 |  |
| Manhattan | New York County | 36061 | 1,628,706 | 22.79 | 71,466 | 9.18 |  |
| Queens | Queens County | 36081 | 2,253,858 | 109.25 | 20,630 | 8.93 |  |
| Staten Island | Richmond County | 36085 | 476,143 | 58.15 | 8,188 | 8.88 |  |
| aggregate | New York City | 99999 | 8,336,817 | 302.06 | 27,600 | 8.90 | 40.7420, 73.9073 |

Data providers/datasets to be used:

1. [co-est2019-alldata.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/co-est2019-alldata_exp.csv), U.S. Census Bureau, “Annual Estimates of the Resident Population for Counties: April 1, 2010 to July 1, 2019.”  
   **3142** records: FIPS **1001–56045** (Alabama–Wyoming), we need:  
   **FIPS | STNAME | CTYNAME | POPESTIMATE2019**  
   [more information here](https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html)
2. [us-counties.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/us-counties.csv), cumulative COVID-19 case and fatality data per county, reported by the *New York Times*

Each day, the *NYT* has appended to this list the updated cumulative per-county count of cases and deaths from COVID, starting with one case in one county in Washington state on January 21, 2020. It continues to be updated daily. When I downloaded it, it was 1.1 million records long. If we wanted more work to do, we could iterate over the list and generate one heat map a day, then import those into Adobe Premiere and make an animation of them. It would work a computer hard but be fairly straightforward to do. In the event, we decided to focus on the cumulative COVID case and death count per county on July 14, 2020, the last day of the *New York Times*/Dynata survey. How many cases and deaths had there been up until then? So much of the pandemic has happened since then, it’s possible people would answer the same survey differently today. We created a CSV with all the data from only July 14, 2020:  
**3136** records: FIPS **1001–56045** (Alabama–Wyoming) plus an aggregate value for all of New York City we’re informally calling FIPS 99999 (the five counties that make up New York City are blank in this listing), as well as 24 additional records that don’t have a FIPS code: one record each for Joplin, Missouri and Kansas City, Missouri, and 22 records reporting cases and deaths associated with a state but not with a county. It’s possible the people in these cases and deaths represent part of the homeless population in that state ([more information here](https://github.com/nytimes/covid-19-data/blob/master/README.md)):  
**date | county | state | fips | cases | deaths**

1. [counties-by-land-area.csv](https://gist.github.com/palewire/5cf017f21730ebd8303fb51e0cc7a2cd), U.S. counties land area by square miles.   
   ArcGIS Hub  
   **3220** records: FIPS **1001–72153** (Alabama–Puerto Rico); we need:  
   **FIPS | STATE\_NAME | NAME [county] | SQMI**

This has a ton more information: count of households, age by strata, male/female, race, acreage in crops (that’d be interesting alongside population density)--if we *wanted* more work to do, we could go deep. ([more information here](https://hub.arcgis.com/datasets/48f9af87daa241c4b267c5931ad3b226_0/data?orderBy=FIPS))

1. county\_centers.csv, geographical centers of each county in the U.S., assembled by B.T. Skinner on GitHub. Matt spot-checked some of these and found them to be accurate.  
   3147 records: FIPS **1001–56045** (Alabama–Wyoming)  
   **fips | clon10 | clat10**

([more information here](https://github.com/btskinner/spatial/blob/master/data/county_centers.csv))

1. [mask-use-by-county.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/mask-use-by-county-exp.csv), results of the *New York Times*/Dynata survey July 2-14, 2020  
   **3142** records: FIPS **1001–56045** (Alabama–Wyoming)  
   **COUNTYFP | NEVER | RARELY | SOMETIMES | FREQUENTLY | ALWAYS**  
   ([more information here](https://github.com/nytimes/covid-19-data/blob/bde13b021e99c6b4a63fb66a6144e889cc635e31/mask-use/README.md))

Steps:

1. Make list of column/headers for *Great Big Dataframe (GBDf)* with empty cells - Paul
   1. **FIPS**
   2. **County**
   3. **State**
   4. **2019PopEst** (2019 population estimate)
   5. **PopDens** (Population density)
   6. **MskScore** (Mask-wearing score)
   7. **Income** (Mean per-capita income; bonus)
   8. **CaseRate** (cumulative COVID cases per 100,000 population by county on July 14, 2020)
   9. **DeathRate** (cumulative COVID deaths per 100,000 population by county on July 14, 2020)
2. Write a routine that can be applied to each of the four source datasets that converts the value that represents FIPS code into string, checks its length with len(), then prepends a 0 to any four-digit strings. **Emerson**
3. Import CSVs
   1. [Co-est2019-alldata.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/co-est2019-alldata_exp.csv) **Aleena**
      1. import into *Census* Dataframe
      2. Conform FIPS
      3. Append 2019 pop. est. to the *GBDf*
   2. [Us-counties.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/us-counties.csv) **Matt**
      1. Import CSV into *CasesAndDeaths* Dataframe
      2. isolate values where date = July 14, 2020
      3. Conform FIPS
      4. Import 2019 pop. est. from *Census* Dataframe
      5. Calculate cases per 100,000 population
      6. Calculate deaths per 100,000 population
      7. Append cases/100,000 pop. and deaths/100,000 pop. to *GBDf*
   3. [Mask-use-by-county.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/mask-use-by-county-exp.csv) **Emerson**
      1. Import into *MaskUse* Dataframe
      2. Conform FIPS
      3. Calculate overall mask-wearing score
      4. Append mask-wearing score to *GBDf*
      5. Merge DataFrames into *GBDf*
   4. [Counties-by-land-area.csv](https://gist.github.com/palewire/5cf017f21730ebd8303fb51e0cc7a2cd) **Paul**
      1. Import into *Census* Dataframe
      2. Conform FIPS
      3. Calculate population density
      4. Append population density to *GBDf*
4. Calculate Pearson’s correlation and create scatter plots with regression analysis:
   1. **Mask-wearing score vs COVID cases and**
   2. **Mask-wearing score vs deaths per 100,000 pop.**
   3. Population density vs COVID cases per 100,000 pop.
   4. Population density vs COVID deaths per 100,000 pop.
   5. (bonus) Mean per capita income vs COVID cases per 100,000 pop.
   6. (bonus) Mean per capita income vs COVID deaths per 100,000 pop.
5. Create heatmaps of:
   1. **Mask-wearing score**
   2. **Cases per 100,000 population**
   3. Deaths per 100,000 population
   4. Population density
   5. Mean per capita income
6. Create PowerPoint presentation with slides: (optional)
   1. Intro
   2. Overview of the “questions asked of our data”
   3. Datasets we used
   4. Specific questions and our results (with visual representations)
   5. Conclusion

Rough breakdown of tasks:

1. Aleena:
   1. [Co-est2019-alldata.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/co-est2019-alldata_exp.csv):
      1. import into *Census* Dataframe
      2. Append 2019 pop. est. to the *GBDf*
2. Emerson:

[Mask-use-by-county.csv](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/Resources/mask-use-by-county-exp.csv):

* 1. Import into *MaskUse* Dataframe
  2. Calculate overall mask-wearing score
  3. Append mask-wearing score to *GBDf*

1. Paul:
   1. [Describe .CSVs](https://github.com/emersonmolett/Project-1/blob/paulbernhardt/README.md)
   2. [Push them to GitHub](https://github.com/emersonmolett/Project-1/tree/paulbernhardt/Resources)
2. Matt:
   1. Import CSV into *CasesAndDeaths* Dataframe
   2. Isolate values where date = July 14, 2020
   3. Import 2019 pop. est. from *Census* Dataframe
   4. Calculate cases per 100,000 population
   5. Calculate deaths per 100,000 population
   6. Append cases/100,000 pop. and deaths/100,000 pop. to *GBDf*
   7. Create heatmaps for items d.) and e.)