**Final Project**

**Comparison of Socio-Economic Factors (Housing Assistance, Education, Unemployment, and Gun Violence) in the United States**

1. **DATA**

1.1 Data Sources, Collection and Limitations

All data sets used are publicly available and were sourced through [Kaggle.com](https://www.kaggle.com/datasets)

This project used multiple data sets, including:

1. [HUD Multifamily Assisted Housing Properties](https://www.kaggle.com/datasets/thedevastator/hud-multifamily-housing-assisted-properties-data)

The data set was compiled by [Mathew Schnars](https://data.world/mschnars) using open administrative data from the U.S. Department of Housing and Urban Development (HUD). It lists all multifamily (five family dwelling units or more) rental housing properties financially supported by HUD to promote and protect affordable rental housing for low-income populations or those with special needs. A noted limitation of the data is that the physical location data for each property is not 100% accurate, which is why I have aggregated to the county level using the Country Federal Information Processing Standards (FIPS) Codes included in the data set as a more reliable geolocator.

1. [Gun Violence Data](https://www.kaggle.com/datasets/jameslko/gun-violence-data)

This data set was made available as a potential source in the Achievement A6 Project Brief with the recognized owern of James Ko and lists data for all recorded incidents of gun violence between January 2013 and March 2018 in the U.S. The data set was created by webscraping of the [Gun Violence Archive (GVA)](http://gunviolencearchive.org/), a not-for-profit organization that provides free online access to gun-related violence which has been collected and checked for accuracy. As a potential limitation, It is not administrative data but is collected by a third-party organization, which is credible. I thought it would be an interesting once to include in the analysis.

1. [US Unemployment & Education Level](https://www.kaggle.com/datasets/valbauman/student-engagement-online-learning-supplement)

The unemployment and education level data sets use administrative data and were drawn from the [County-level Data Sets](https://www.ers.usda.gov/data-products/county-level-data-sets/) from the USDA Economic Research Service at the U.S. Department of Agriculture. The data sets were created for use in the [LearnPlatform COVID-19 Impact on Digital Learning](https://www.kaggle.com/c/learnplatform-covid19-impact-on-digital-learning) Kaggle competition. I selected these data sets as they have locations mapped to the State and County FIPS codes, which would allow me to merge them with the HUD data set. The data set also included an excel tool to map State and County names with ZIP and State and County FIPS codes, which I also will use to merge the data sets by editing it and converting to a .csv file (State\_County\_FIPS.csv).

1. [US Census State and Country Shapefiles](https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html) (to map longitude and latitude to State and County FIPS codes)

I downloaded these shape files from the U.S. Census Bureau as open, administrative data sets from the Census Bureau’s MAF/TIGER geographic database. They are boundary files that are designed for small scale thematic mapping. They were the source to convert the latitude and longitude geolocators from the Gun Violence Data data set to State and Country FICS codes using Jupyter notebooks, so the data set could be merged with the other four data sets which already contained the FICS codes.

Data Set Selection

I selected these data sets as I was interested to compare different socio-economic factors and investigate for any relationships or correlations that would yield insights. I was particularly interested in looking at social housing support and after searching data sets for different countries I settled on HUD data as there were rich sources of other data sets that could be geolocated with the HUD data. After importing and investigating several data sets in Jupyter notebooks and Excel, I settled on four data sets that could be geolocated at the county level (with some additional wrangling of the Gun Violence Data data set).

1.2 Data Set Profiles

1.2.1 Data Set Cleaning

a) HUD data set

* The original data set had 286 columns and 23,634 rows.
* I started the cleaning process by removing columns that were not relevant to the planned analysis. Further data wrangling of the other data sets showed that additional columns were not required.
* I renamed the remaining column variables for ease of understanding (variable ‘X’ was renamed ‘Longitude” for example, before it was eventually removed).
* Data consistency checks showed the presence of 1,433 missing values, which were removed.
* After further column reductions after standardizing geolocation information for State, Count, and FIP codes from another data set this will be merged with the HUD data set was left with 2 columns and 23,598 rows

b) Gun Violence Data data set

* The original data set had 28 columns and 239,677 rows
* I started the cleaning process by removing columns that were not relevant to the planned analysis. Further data wrangling of the other data sets showed that additional columns were not required.
* Data consistency checks showed the presence of 38,772 and 32,335 missing values from two variables (house and senate representational boundaries) I was exploring using – these were dropped. Two other variables had 16,497 and 11, 944 missing variables (congressional representational boundaries and physical street addresses), which were also dropped. Additional missing values were identified in the latitude and longitude values with 7,923 missing values each. These represented 3% of the total data set, and because they could not be inputed from the other geolocational data available in the data set, I removed these.
* After further column reductions after standardizing geolocation information for State, Count, and FIP codes from another data set this will be merged with the Gun Violence data set was left with 3 columns and 231,754 rows

c) US Unemployment and Education Level data set

* The original education data set had 48 columns and 3,283 rows
* I started the cleaning by dropping columns that were not required – primarily variables with absolute values, which were not needed for the analysis.
* Data consistency checks identified 197 missing values. There was no reliable method to impute these values, which accounted for <1.5% of the total data set, so they were removed.
* After further column reductions after standardizing geolocation information for State, Count, and FIP codes from another data set this will be merged with the Education data set was left with 17 columns and 3,260 rows
* The original unemployment data set had 93 columns and 3,275 rows
* I started the cleaning by dropping columns that were not required – again, primarily variables with absolute values, which were not needed for the analysis.
* Data consistency checks identified 341 missing values. While missing unemployment rate values could be imputed using linear analysis, given the majority came the same year and represented 2.5% of the total data set, I decided to remove the missing values.
* After further column reductions after standardizing geolocation information for State, Count, and FIP codes from another data set this will be merged with the Unemployment data set was left with 24 columns and 3,180 rows.

1.2.2 Data Profiles

a) cleaned HUD data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Column | Description | Time Variant/ Invariant | Data Type |
| 1. | Property Index # | Index #, unique for each multifamily property in the HUD data set | Invariant | Quantitative, continuous |
| 2. | Has Active Financing | Does the multifamily property continue to receive HUD financing assistance | Variant | Qualitative, binary |
| 3. | FIPS Code | standardized State and County location code | Invariant | Qualitative, ordinal |

\* County name and State Name will be added to the FIPS Code

b) cleaned Gun Violence data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Column | Description | Time Variant/ Invariant | Data Type |
| 1. | date | Index #, unique for each multifamily property in the HUD data set | Invariant | Quantitative, continuous |
| 2. | gun incident ID | Does the multifamily property continue to receive HUD financing assistance | Variant | Qualitative, nominal |
| 3. | FIPS Code | standardized State and County location code | Invariant | Qualitative, ordinal |

\* County name and State Name will be added to the FIPS Code

c) cleaned Education data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Column | Description | Time Variant/ Invariant | Data Type |
| 1. | FIPS Code | standardized State and County location code | Invariant | Qualitative, ordinal |
| 2. | Percent of adults with less than a high school diploma, 1980 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 3. | Percent of adults with a high school diploma only, 1980 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 4. | Percent of adults completing some college (1-3 years), 1980 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 5. | Percent of adults completing four years of college or higher, 1980 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 6. | Percent of adults with less than a high school diploma, 1990 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 7. | Percent of adults with a high school diploma only, 1990 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 8. | Percent of adults completing some college or associate's degree, 1990 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 9. | Percent of adults with a bachelor's degree or higher, 1990 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 10. | Percent of adults with less than a high school diploma, 2000 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 11. | Percent of adults with a high school diploma only, 2000 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 12. | Percent of adults completing some college or associate's degree, 2000 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 13. | Percent of adults with a bachelor's degree or higher, 2000 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 14. | Percent of adults with less than a high school diploma, 2015-19 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 15. | Percent of adults with a high school diploma only, 2015-19 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 16. | Percent of adults completing some college or associate's degree, 2015-19 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |
| 17. | Percent of adults with a bachelor's degree or higher, 2015-19 | % of adults in the county at that level of formal education | Time Variant | Quantitative, continuous |

\* County name and State Name will be added to the FIPS Code

d) cleaned Unemployment data set

2 3180 non-null float64

3 3180 non-null float64

4 3180 non-null float64

5 3180 non-null float64

6 3180 non-null float64

7 3180 non-null float64

8 3180 non-null float64

9 3180 non-null float64

10 3180 non-null float64

11 3180 non-null float64

12 3180 non-null float64

13 3180 non-null float64

14 3180 non-null float64

15 3180 non-null float64

16 3180 non-null float64

17 3180 non-null float64

18 3180 non-null float64

19 3180 non-null float64

20 3180 non-null float64

21 3180 non-null float64

22 3180 non-null int8

23 3180 non-null float64

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Column | Description | Time Variant/ Invariant | Data Type |
| 1. | FIPS Code | standardized State and County location code | Invariant | Qualitative, ordinal |
| 2. | Unemployment\_rate\_2000 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 3. | Unemployment\_rate\_2001 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 4. | Unemployment\_rate\_2002 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 5. | Unemployment\_rate\_2003 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 6. | Unemployment\_rate\_2004 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 7. | Unemployment\_rate\_2005 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 8. | Unemployment\_rate\_2006 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 9. | Unemployment\_rate\_2007 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 10. | Unemployment\_rate\_2008 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 11. | Unemployment\_rate\_2009 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 12. | Unemployment\_rate\_2010 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 13. | Unemployment\_rate\_2011 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 14. | Unemployment\_rate\_2012 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 15. | Unemployment\_rate\_2013 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 16. | Unemployment\_rate\_2014 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
| 17. | Unemployment\_rate\_2015 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Unemployment\_rate\_2016 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Unemployment\_rate\_2017 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Unemployment\_rate\_2018 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Unemployment\_rate\_2019 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Unemployment\_rate\_2020 | Rate of unemployment in the county for that year | Time Variant | Quantitative, continuous |
|  | Median\_Household\_Income\_2019 | Median HH income in that county | Time Variant | Quantitative, continuous |
|  | Med\_HH\_Income\_Percent\_of\_State\_Total\_2019 | Median HH income in the county as a percentage of the Median HH income of the State | Time Variant | Quantitative, continuous |

\* County name and State Name will be added to the FIPS Code

\*\* I tried to merge the HUD\_Gun Violence dataframe with the State\_County\_FIPS csv file to add standard State and County names to each FIPS code, but each time I tried to do this merge I received a ‘kernel died’ error. Is there anything I can do to try and merge these data sets into one dataframe?

1.3 Data Questions

There are many potential questions to explore with these four data sets

1. Is there any trend/ difference in gun violence, educational attainment, and/ or unemployment rates for areas that continued to receive multifamily housing assistance and those no longer receive multifamily housing assistance?
2. Is there any relationship with areas that receive multifamily housing assistance and gun violence?
3. Is there any relationship with areas that receive multifamily housing assistance and education attainment?
4. Is there any relationship with areas that receive multifamily housing assistance and unemployment?
5. Is there any relationship between areas with gun violence and educational attainment?
6. Is there any relationship between areas with gun violence and unemployment?
7. Is there any relationship between educational attainment and unemployment?

Note – I had looked for data sets showing rental rates for multifamily properties, but I was not able to match data sets I found, such as from Zillow, with county/ FIPS-level location data, but I will continue to investigate as I continue with this project to try and include this as well.