Socioeconomic Analysis of Maryland Counties

IBM Data Science Specialization

This report was completed in partial requirement for completion of the IBM Applied Data Science Capstone, which is Course 9 of 9 in the IBM Data Science Specialization hosted by Coursera.

The complete notebook can be viewed at:
https://nbviewer.jupyter.org/gist/yrousselet/cea94f0580e333c75a791330d48c1339

Contents

Introduction

Maryland, the 7th State to join the Union in 1788, is the 19th most populous state in the union [1]. Part of the Mid-Atlantic region, it has common boarders with 4 states, Virginia, West Virginia, Pennsylvania, and Delaware, as well as the District of Columbia [2]. While its largest and most well-known city is Baltimore, its capital is Annapolis, located on the Chesapeake Bay at the mouth of the Severn River, roughly 25 miles south of Baltimore [3].

After finishing graduate school, I was hired by a Maryland-based company, and moved from Southern California to the Greater Baltimore Area. More than five years after moving to Maryland, I remain unfamiliar with this state. Therefore, I have decided that the focus of this project will be the great State of Maryland. More specifically, while Maryland is the richest State as defined by median household income, and second richest territory after Washington D.C. [4], I am interested in finding out the demographic and wealth distributions, among other data, at a county level, to better understand whether or not every county fares as well as the state level metrics would indicate.

As mentioned, my approach will be to focus on data at the county level. After pre-processing and exploring the selected datasets, I will use advanced data analytics techniques to gain additional insights into the make-up of the Maryland counties. This project is a great opportunity to demonstrate and further refine my newly acquired Data Science skills while learning something new along the way, and I hope that readers interested in deepening their understand and knowledge of the State of Maryland and its counties will find it instructive and beneficial.

Data

Data Sources

The State of Maryland, through its Open Data Portal, provides free and unlimited access to geographic and non-geographic datasets for search and exploration. This Portal is managed by the Maryland Department of Information Technology, with contributions from many State and Local agencies [5]. Most of the datasets for this project will be obtained from this portal.

The objective is to analyze datasets at the county level, with a focus on demographic and economic measures. To do so, the following datasets will be used:

- Maryland Counties Socioeconomic Characteristics [6]
- Choose Maryland: Compare Counties - Demographics [7]
- Choose Maryland: Compare Counties - Quality of Life [8]
- Choose Maryland: Compare Counties - Education [9]
- Choose Maryland: Compare Counties - Workforce [10]
- Violent Crime & Property Crime by County: 1975 to Present [11]

Additionally, geographical data detailing the characteristics [12] and the boundaries [13] of the Maryland counties will be used for data visualization purposes.

Finally, additional data might be obtained from additional sources as needed. Appropriate references will be included when necessary. No Foursquare location data will be included, as authorized by a member of the Coursera Teaching Staff in a recent reply to Discussion Forum post [14].

Note 1: In order to simplify the data processing and reduce complexity and compute time requirements, all datasets have been exported to either CSV or GeoJSON formats.

Note 2: Most of the available data is current as of 2016; while it would be best to have the most up to date datasets to work with, for the context of this work "mismatched" datasets dates, and 3+ year-old datasets are considered satisfactory

Data Cleaning and Feature Selection

The first step of this process is to create a reference dataframe which contains select data from the datasets listed above. County-level datasets focused on education, demographics, workforce, quality of life, and crime, along with datasets containing information related to county seats and their location, was consolidated in a reference dataframe. The final dataframe has 24 rows (1 per county), and 23 columns (or features). Table 1 shows the 23 features of this dataset, and the first five rows of data.

Table 1.a. Part of reference dataframe showing 12 of 23 features, and 5 rows of data

	Jurisdictions	County Seat	Area (Square Miles)	Latitude	Longitude	Total Population	Population Density (People per Square Mile)	Total Households	Median Age (Years)	Median Household Income ($)	Number of 2 and 4-Year Colleges and Universities per 100,000 People of Voting Age	Population 25 years and older without a High School Diploma (%)
0	Allegany County	Cumberland	430.0	39.652650	-78.762383	70975.0	165.0	27759.0	41.4	43535.0	3.5	10.5
1	Anne Arundel County	Annapolis	588.0	38.978640	-76.492786	576031.0	980.0	205395.0	38.4	97051.0	0.7	8.0
2	Baltimore City	Baltimore City	92.0	39.290882	-76.610759	602495.0	6549.0	239791.0	35.6	50501.0	2.7	15.8
3	Baltimore County	Towson	682.0	39.401855	-76.602388	828431.0	1215.0	312859.0	39.5	75836.0	1.3	8.9
4	Calvert County	Prince Frederick	345.0	38.540554	-76.583507	92003.0	267.0	31462.0	40.5	106270.0	0.0	6.2

Table 1.b. Part of reference dataframe showing 11 of 23 features, and 5 rows of data

Population 25 years and older with Bachelor's Degree or Higher (%)	Median Sale Price of a Home ($)	Ratio of Median Housing Unit Sales Price to Median Household Income	Number of Housing Units Sold per 100000 Households	Number of Business Establishments per 100,000 People of Voting Age	Number of Business Establishments with 100 or More Workers per 100,000 People of Voting Age	Unemployment Rate (%)	Cost of Living Index	Percentage of Families Living in Poverty	Percentage of Housing Units that are Vacant	Crime Rate per 100,000 People
18.2	90829.0	2.1	2247.9	2590.3	43.4	5.5	84.9	10.6	16.0	3467.1
40.1	338287.0	3.5	4196.3	3507.6	90.4	3.3	119.6	3.9	6.9	2450.1
30.4	139723.0	2.8	3523.9	2953.5	87.9	5.7	101.3	17.2	18.7	7033.7
37.8	238426.0	3.1	3191.9	3456.6	92.9	4.0	109.9	6.0	7.0	3402.1
30.1	318471.0	3.0	5009.2	2808.0	28.9	3.5	121.7	3.3	9.6	1155.2

The data gathering stage of this work is now complete. This data will be used in the next steps of this analysis, to gain more insights on the make-up of the 24 counties in Maryland.

Methodology

As discussed in the introduction, the purpose of this work is to study the make-up of the various counties in Maryland, and to gain more insights into how similar and/or different from each other. Now that we have collected geographical, socio-economic and demographics data about the 24 Maryland counties, and compiled it into a reference dataframe, we will perform basic exploratory analysis to gain some high level information about the counties, mostly using cartographical and heat maps type visualizations. To do so, we will use the following packages/libraries:

- Geopy to find geographical coordinates
- Folium to generate maps
- Seaborn and Matplotlib to plot data

Finally, we will cluster the counties using a subset of the available data, to find which counties are similar to each other, and which are not. To do so will use the scikit-learn Machine Learning library, and more specifically the k-means clustering algorithm. Detailed data for each cluster will then be used to understand their main characteristics. Hopefully, these clusters will provide enough insights into the data which can be used as a starting point to further ones understand of the similarities and differences between each county, and possibly inform relocation of investment decisions for prospective buyers/investors.

Exploratory Data Analysis

First, let's create a map of Maryland with markers representing the council seats, using the Folium package for generating the map, and geopy to obtain geographical coordinates. To get a better understanding of the population centers, we can generate a Choropleth map with the counties shaded in colors as a function of total population, with tooltips providing information on key socioeconomic data.

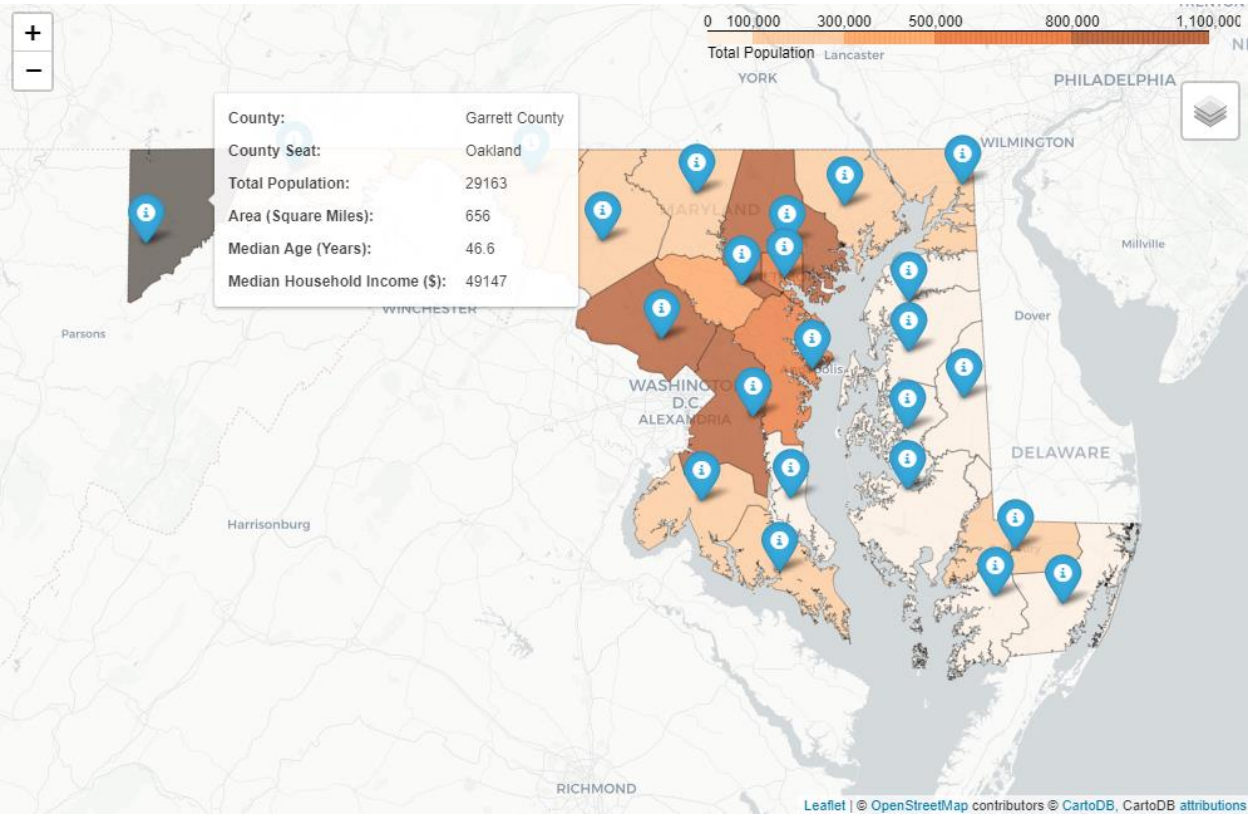

Figure 1. Chloropleth map of Maryland counties shaded by total population, with markers indicating county seats

Figure 1 shows that there are two major population centers in Maryland, Baltimore City and its suburbs, and the suburban area north and east of Washington DC. The remaining counties are sparsely populated.

A good starting point for exploring the data is evaluate the correlation between each parameter. To do so, we can use the Pearson Correlation Coefficient, and heat maps to visualize the results. County and county seat names, longitude and latitude, and absolute population and area metrics are removed to simplify/shrink the dataset.

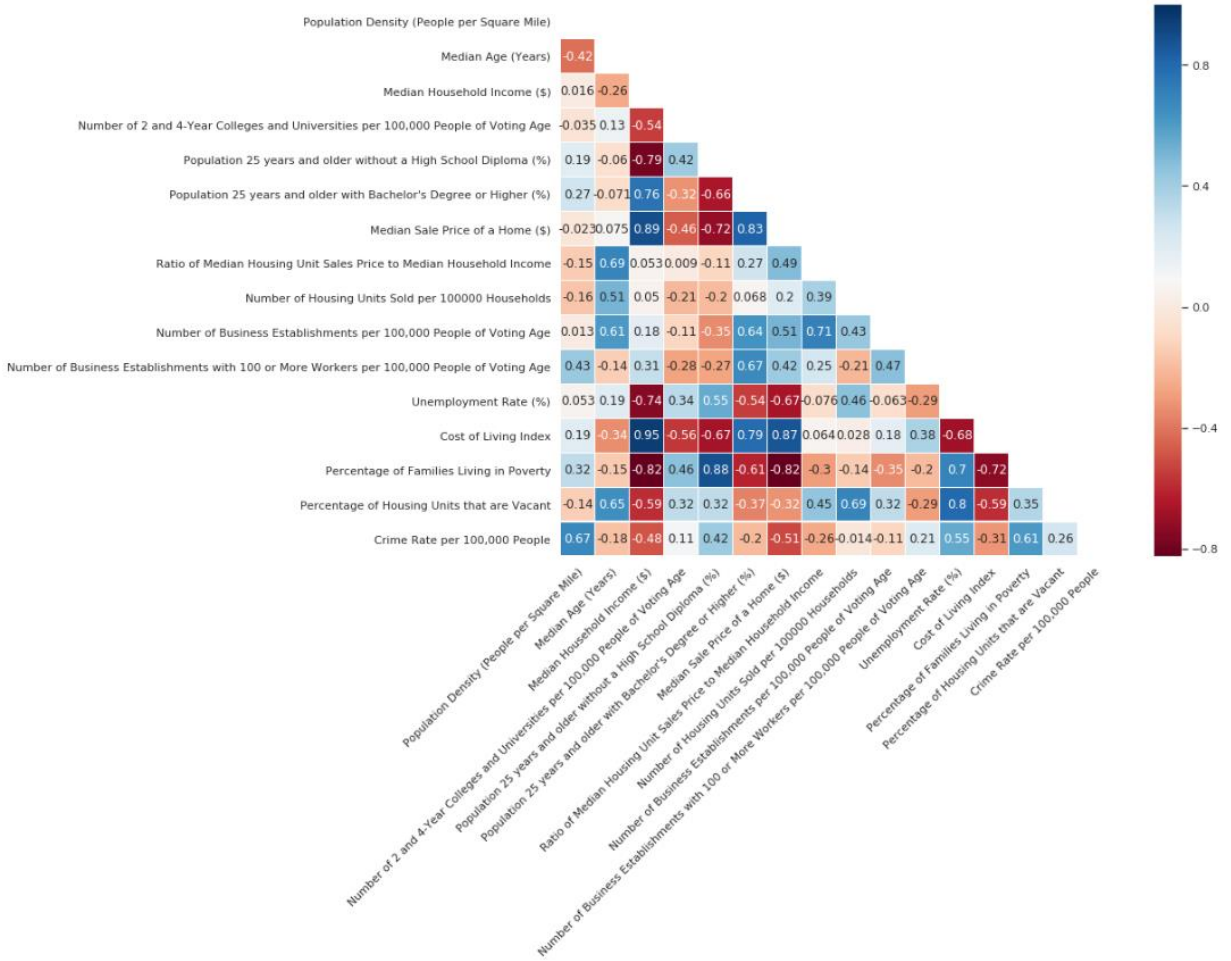

Figure 2. Heatmap of Pearson Correlation Coefficients for select features of the reference dataframe

From Figure 2, we can easily identify "pockets" of highly correlated parameters, such as education/employment metrics, wealth/housing metrics, and poverty/employment/crime metrics. We can use these insights to create reduced datasets to further explore the data. We will create the following 4 "reduced" datasets:

- General
- Education and Employment
- Housing
- Hardship

General

Includes population, population density, median age, and median income, sorted in descending order by population.

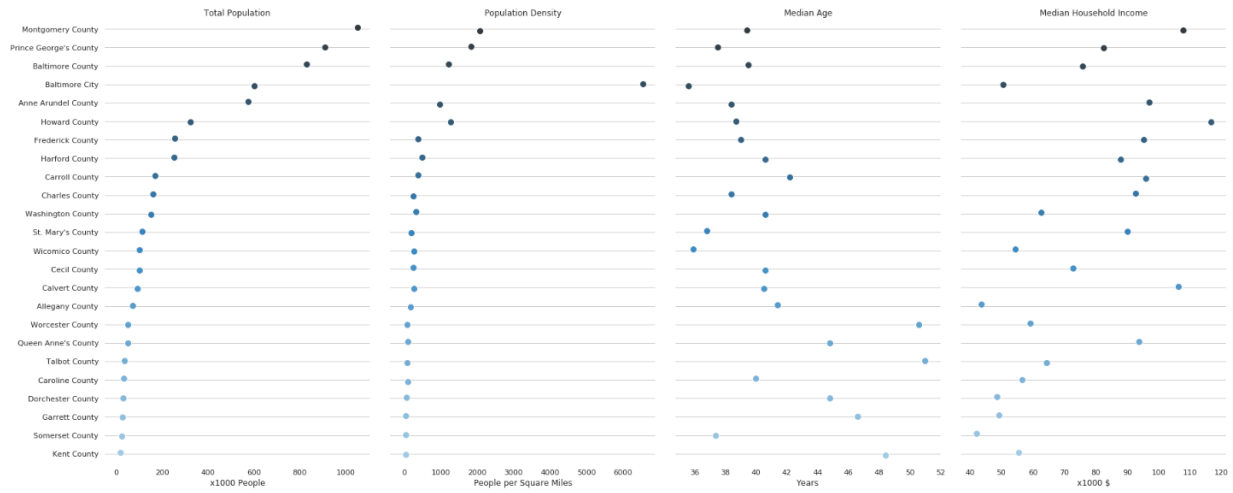

Figure 3. Population, population density, median age, and median income per county, sorted in descending order of population

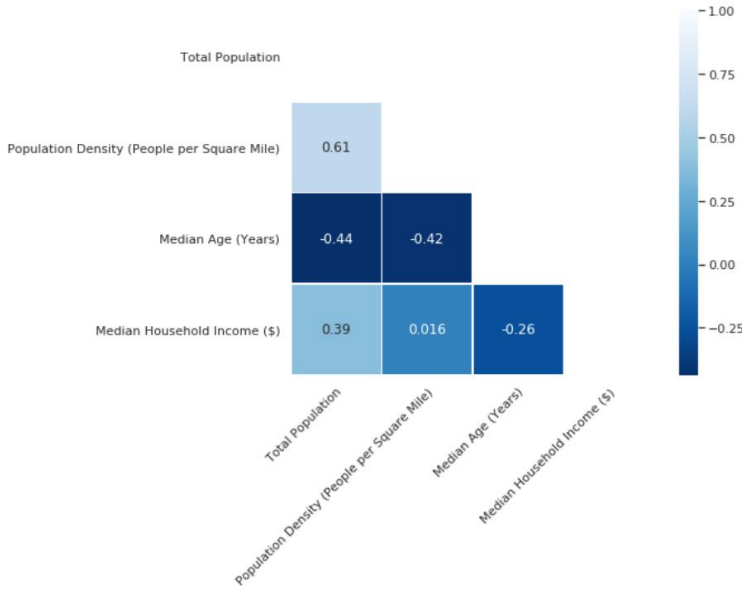

Figure 4. Heatmap of Pearson Correlation Coefficients for the features plotted in Figure 3

Figures 3 and 4 show that the difference between the make-up of the 24 counties is striking. Looking at population alone, there is a roughly 10x difference in population between the most populated county, Montgomery, and the 10 least populated ones. Additionally, the data shows that for these 4 metrics, the difference between counties are significant, and it is hard to draw any conclusions, beyond that it appears that the most populated counties "skew" younger and "richer". Finally, note that Baltimore City's population density is much higher than that of the other counties.

Education and Employment

Includes the median household income, the percentage of population 25 years and older without a high school diploma, the percentage of population 25 years and older with bachelor's degree or higher, and the unemployment rate.

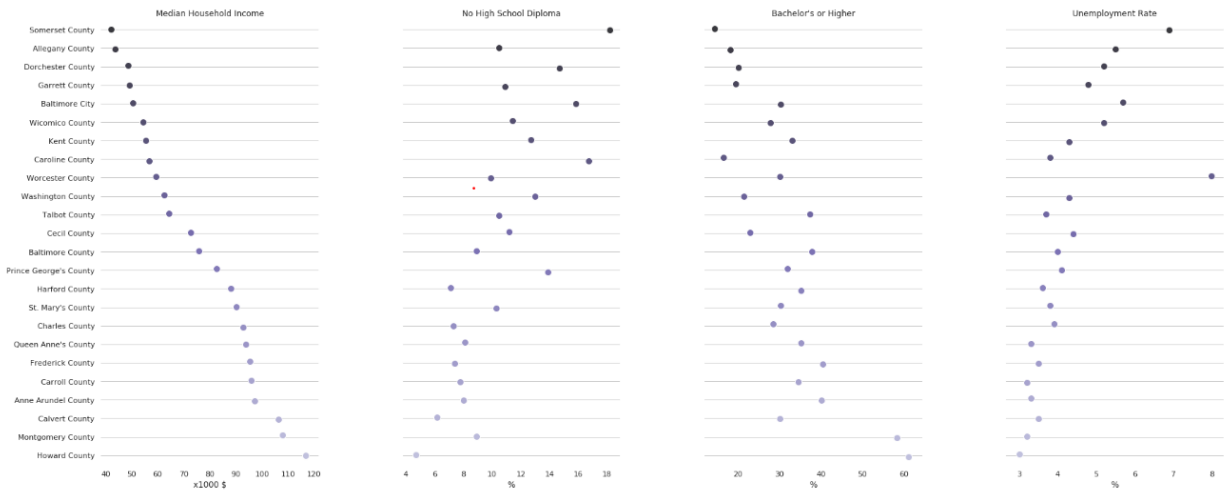

Figure 5. Median household income, population 25 years and older without a high school diploma, population 25 years and older with bachelor's degree or higher, and unemployment rate per county, sorted in descending order of household income

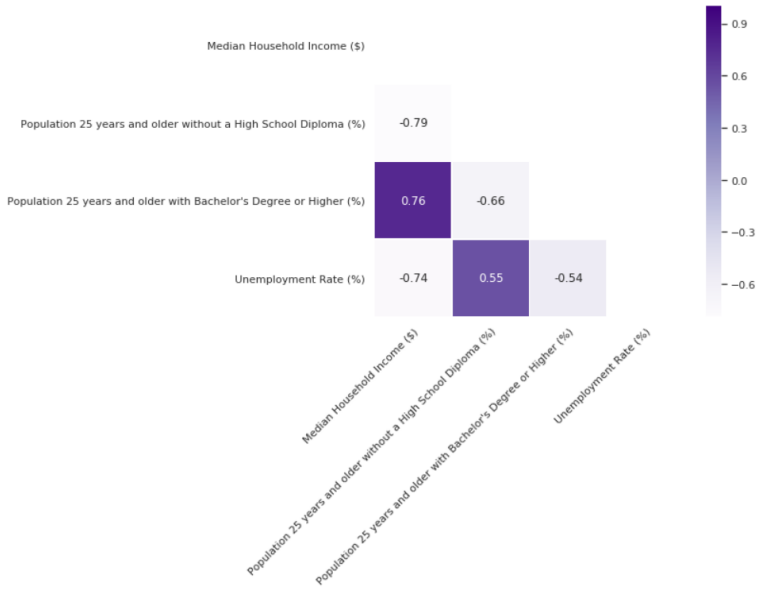

Figure 6. Heatmap of Pearson Correlation Coefficients for the features plotted in Figure 5

As for Figures 3 and 4, Figures 5 and 6 the difference between counties is quite noticeable; Figure 5 shos that there is a twofold difference in household income between the 3 "richest" counties, and the 10 poorest ones. Figure 6 shows that household income is strongly correlated to educational achievement and unemployment, and that the percentage of population over 25 years old without a high school degree and the percentage of this same population with a bachelors or higher are negatively correlated. Note that Worcester County's unemployment rate is much greater that any other county, especially when compared to median household income.

Housing

Includes the median sales price of a home, the cost of living index, the ratio of median home sales price to median household income, and the number of vacant units.

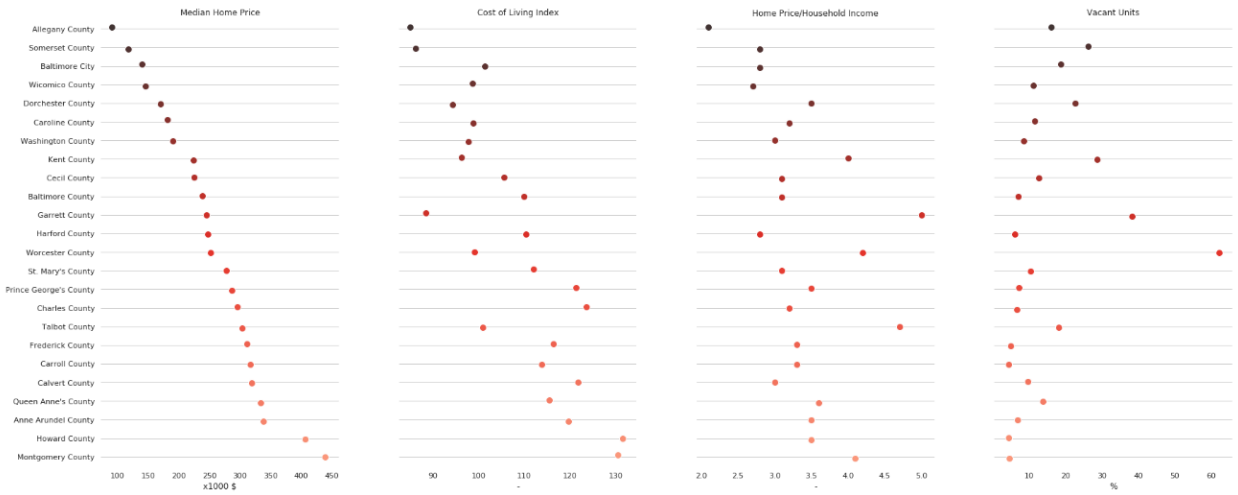

Figure 7. Median home sales price, cost of living index, ratio of median home sales price to median household income, and number of vacant units per county, sorted in descending order of median home sales price

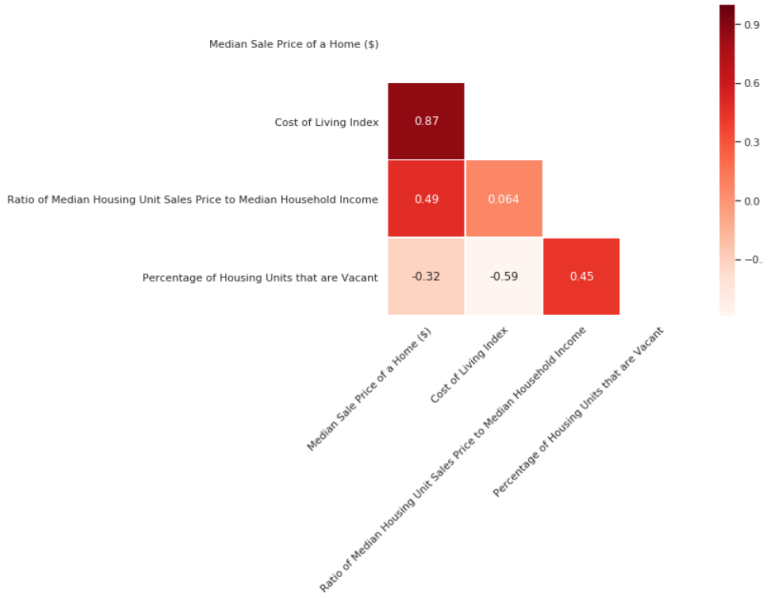

Figure 8. Heatmap of Pearson Correlation Coefficients for the features plotted in Figure 7

Figure 7 shows that the difference in median housing unit sale price between counties is impressive, and as expected is strongly correlated to cost of living, as showed in Figure 8. As expected, median housing price is greater in sought-after counties, i.e. counties that have very few vacant housing units, as can be seen in Figure 8 by the negative correlation coefficient. The ratio of median home price to median household income is very consistent between counties, and outside of a few outliers is generally between 3 and 4.

Hardship

Includes the unemployment rate, the percentage of the population 25 years and older without a college degree, the percentage of people living in poverty, and the crime rate per 100000 people.

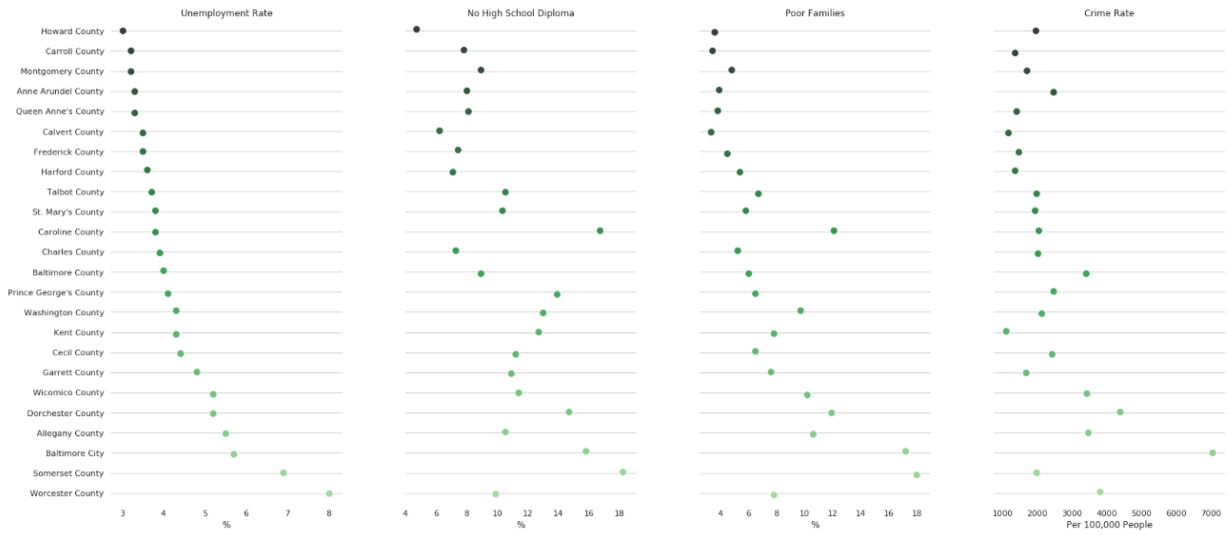

Figure 9. Unemployment rate, population 25 years and older without a college degree, poverty rate, and crime rate per county, sorted in descending order of unemployment rate

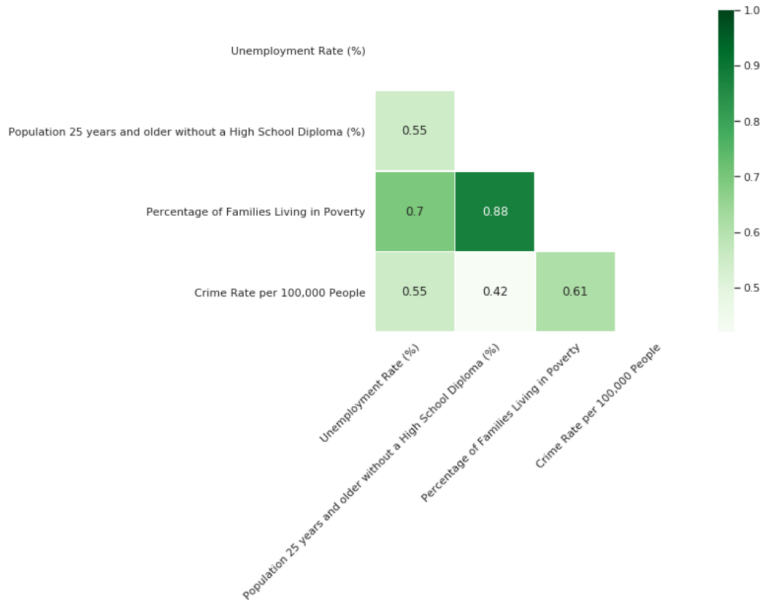

Figure 10. Heatmap of Pearson Correlation Coefficients for the features plotted in Figure 9

As expected based on median household income data, Figure 9 shows that the unemployment and poverty levels vary widely between counties, with some counties faring much better than others. Figure 10 shows a strong correlation between unemployment rate and the percentage of the adult population without a high school degree, as well as with the number of poor families. Crime rate is also positively correlated to unemployment and educational achievement, but the signal in the data is not as strong.

Cluster Counties

We now have gained a better understanding of the similarities and differences between Maryland counties, and where able to visualize this information in easy to understand formats. However, the quantity of data available makes it difficult to be able to compare counties affectively, and thus we will now move to more advanced data analytics to gain more insights into this data.

Similar counties will be clustered together, using the k-means clustering algorithm. The Folium library will be used to visualize the counties in Maryland and their emerging clusters. Based on this clustering, one can easily determine the characteristics of each county.

Create Reference Dataframe

For this analysis to be meaningful, we need to reduce the dataset. The number of samples being low (24 counties = 24 samples), we will limit this clustering to a portion of the available data. Given the large disparity in population size, we will refrain from using absolute metrics, and will use rates, ratios, and other metrics relative to population wherever possible.

We will tackle this analysis by clustering the counties based on metrics that were deemed to be most representative of the various counties based on the exploratory data analysis, based on the exploratory data analysis performed:

- Population Density
- Median Age
- Median Household Income
- Median Home Sales Price
- College Education
- Unemployment Rate
- Poverty Rate
- Crime Rate

The resulting dataframe is shown in Table 2.

Table 2. First 5 rows of dataframe used for clustering

	Jurisdictions	Population Density (People per Square Mile)	Median Age (Years)	Median Household Income ($)	Median Sale Price of a Home ($)	Unemployment Rate (%)	Population 25 years and older with Bachelor's Degree or Higher (%)	Percentage of Families Living in Poverty	Crime Rate per 100,000 People
0	Allegany County	165.0	41.4	43535.0	90829.0	5.5	18.2	10.6	3467.1
1	Anne Arundel County	980.0	38.4	97051.0	338287.0	3.3	40.1	3.9	2450.1
2	Baltimore City	6549.0	35.6	50501.0	139723.0	5.7	30.4	17.2	7033.7
3	Baltimore County	1215.0	39.5	75836.0	238426.0	4.0	37.8	6.0	3402.1
4	Calvert County	267.0	40.5	106270.0	318471.0	3.5	30.1	3.3	1155.2

Find optimal k value

To be able to cluster the counties together, we first need to determine the optimal number of clusters (k value) by building the clustering model and calculating the values of Inertia and Silhouette Coefficient, for k between 1 to 9. We can then use the different values of Inertia and the elbow method, or the Silhouette Coefficients to determine the optimal k value.

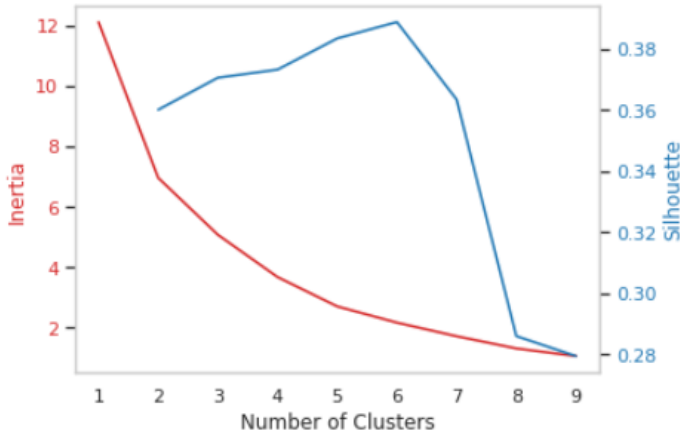

Figure 11. Inertia and Silhouette Coefficient as a function of number of clusters

As can be seen in Figure 11, the elbow method is non-conclusive, and therefore we need to rely on the silhouette coefficients, and settle on k = 6; we will then use this value moving forward

Generate clusters

We then use the k-means algorithm to cluster the counties into 6 clusters. The number of counties per cluster is as follows:

- Cluster 0: 11 counties
- Cluster 1: 6 counties
- Cluster 2: 2 counties
- Cluster 3: 1 county
- Cluster 4: 1 county
- Cluster 5: 3 counties

It appears that a majority of counties fall within the first cluster (label 0). The here are also two “outliers” (i.e. single-county clusters), Baltimore City, and Worcester County. A clustering algorithm such as DBSCAN might have labeled Baltimore City and Worcester County as noise/outliers.

Results and Discussion

In order to analyze these results, we first need to analyze how these clusters were "built". Let's first explore these clusters a bit deeper.

Explore clusters

Let's examine each cluster and determine the discriminating venue categories that distinguish each one. As previously mentioned, there are two "outliers", cluster 3 (Baltimore City), and cluster 4 (Worcester County); additionally, more than 45% of counties are part of cluster 0.

Table 3. Mean values of each feature of the 6 clusters, sorted by median household income in descending order

Cluster Labels	Population Density (People per Square Mile)	Median Age (Years)	Median Household Income ($)	Median Sale Price of a Home ($)	Unemployment Rate (%)	Population 25 years and older with Bachelor's Degree or Higher (%)	Percentage of Families Living in Poverty	Crime Rate per 100,000 People
2	1674.000000	39.050000	112238.500000	422569.000000	3.100000	59.750000	4.200000	1820.300000
0	573.272727	39.845455	89980.272727	289852.636364	3.690909	33.400000	4.936364	1945.954545
4	75.000000	50.600000	59266.000000	251338.000000	8.000000	30.100000	7.800000	3793.200000
5	56.333333	48.666667	56302.000000	257610.000000	4.266667	30.066667	7.366667	1579.500000
1	158.000000	40.016667	51301.833333	149217.333333	5.150000	19.766667	12.083333	2896.383333
3	6549.000000	35.600000	50501.000000	139723.000000	5.700000	30.400000	17.200000	7033.700000

Table 3 paints a clear picture of why the clusters were built as such. More specifically:

- Cluster 0: Upper middle-class suburban counties around Baltimore and along the Baltimore-Washington parkway
- Cluster 1: Rural working-class counties with lower educational achievement, double digit poverty rates, and above average crime rates
- Cluster 2: Densely populated and wealthy DC and Baltimore suburban counties with very high educational achievement, low crime, and above average housing prices
- Cluster 3: Baltimore City; very densely populated, with high crime and poverty rates, and low housing prices and income
- Cluster 4: "Retirement"/older coastal county, with older population, and above average crime rate most likely due to the tourism economy
- Cluster 5: "Retirement"/older rural counties, with older population, low population density, but lowest crime rate and average housing prices

We can also look at standard deviation data to ensure the data within each cluster does not vary significantly.

Table 4. Standard of each feature of the 6 clusters, sorted cluster number in ascending order

Cluster Labels	Population Density (People per Square Mile)	Median Age (Years)	Median Household Income ($)	Median Sale Price of a Home ($)	Unemployment Rate (%)	Population 25 years and older with Bachelor's Degree or Higher (%)	Percentage of Families Living in Poverty	Crime Rate per 100,000 People
0	539.671028	2.268199	9785.790965	38821.755579	0.380669	5.265548	1.210184	689.520390
1	112.550433	3.128205	7949.573156	38852.310174	1.070981	4.667619	3.048552	998.979929
2	568.513852	0.494975	6336.383866	22559.534747	0.141421	2.050610	0.848528	185.120555
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	18.823744	2.212088	7579.782912	41361.257198	0.550757	9.304479	0.585947	443.535309

Standard deviation data in Table 4 shows that with the exception of population density, the standard deviation for each parameter within each cluster is acceptable.

Finally, let's look at the total population per cluster:

- Cluster 0: 3,511,050 residents
- Cluster 1: 416,073 residents
- Cluster 2: 1,375,763 residents

- Cluster 3: 602,495 residents
- Cluster 4: 51,823 residents
- Cluster 5: 85,514 residents

One can see that more than 75% of the state's population is accounted for in clusters 0 and 2. Given that they are the wealthiest counties, one can easily understand what makes Maryland the richest state, using median household income as reference.

Final Visualization

To wrap up this analysis, we can generate an interactive map with the counties shaded based on their clusters, with tooltips providing cluster number and detailed information on key geographical and socioeconomic data.

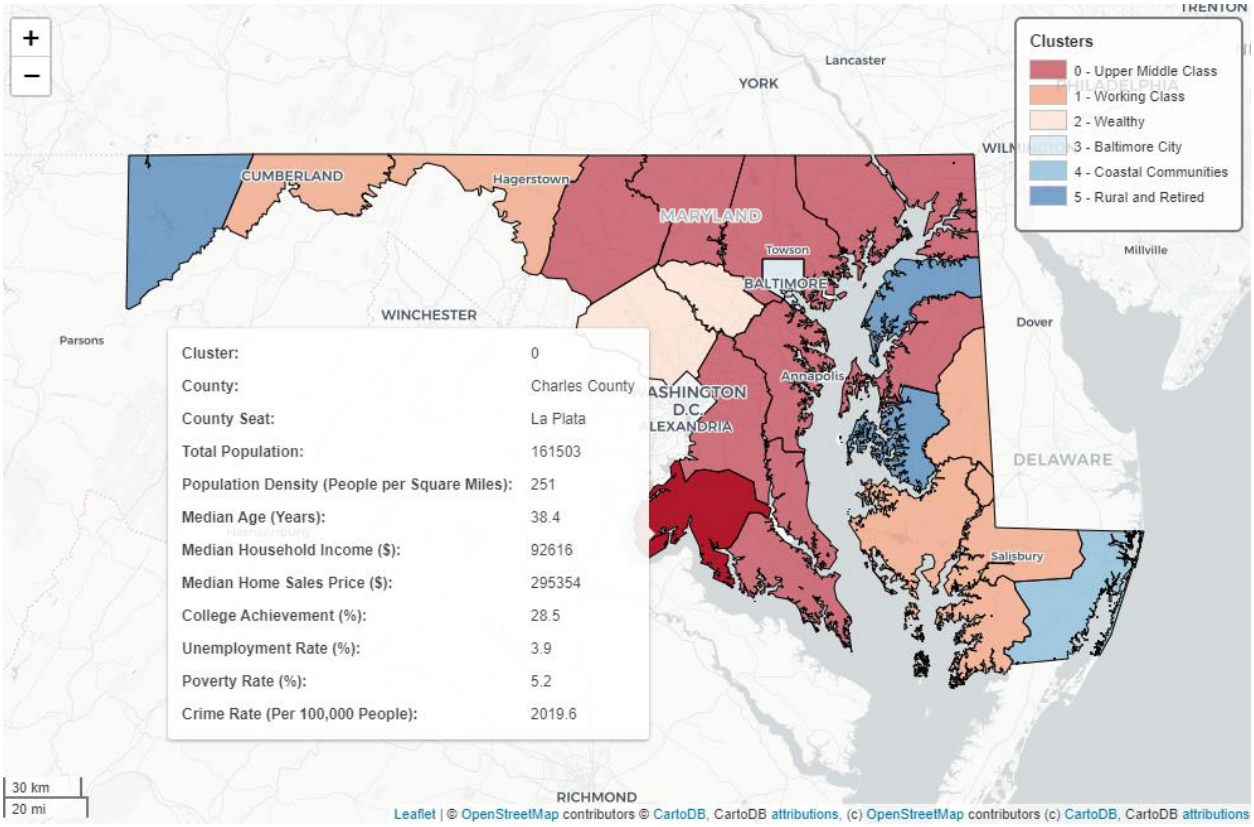

Figure 12. Chloropleth map of Maryland counties shaded by cluster, with “draggable” categorical legend and tooltips

Summary

After passing through numerous datasets, we reduced our analysis to the following key metrics:

- Population Density
- Median Age
- Median Household Income
- Median Home Sales Price
- College Education
- Unemployment Rate
- Poverty Rate
- Crime Rate

Using these metrics, our analysis showed that the Maryland counties can clustered into 6 different groups. Note that 2 of these clusters have only one element, and thus are considered outliers, despite representing over 10% of the State's population. The population in the remaining 4 clusters can be described as follows:

- Suburban, wealthy and highly educated
- Suburban, upper middle class
- Rural, working class
- Rural, older/retired

As mentioned in the introduction, Maryland is the richest State in the Union, using median household income as a metric, and this analysis clearly supports it. Indeed, more than 75% of the population is classified as either wealthy or upper middle class (on a cluster basis). However, this analysis also shows that the struggles of Baltimore City residents are real, and that the opportunities for working class folks to live and work close to population centers might be very limited.

Conclusions

In this workbook, we successfully:

- Compiled and explored data on Maryland Counties
- Sorted the counties and clustered them using the k-means algorithm
- Displayed the clusters and counties on an interactive map

The stated objective of this work was to learn more about the State of Maryland through the make-up of its various counties, and this objective was achieved. The analysis revealed that a large majority of the counties are doing well overall, but the struggle of a minority of the counties, namely Baltimore City and rural, working class counties are real.

More granularity in the data would most likely prove to be more insightful and would paint a picture of inequality even in counties that are deemed well-off. An option could be to repeat a similar analysis using ZIP Codes instead of counties; a quick Google search revealed that Maryland has around 600 ZIP Codes [15], which would provide a very significant increase in sample size and thus a greater ability to discovery more insights in the data.

References

[1] https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population

[2] https://en.wikipedia.org/wiki/Maryland

[3] https://en.wikipedia.org/wiki/Annapolis,_Maryland

[4] https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income

[5] https://doit.maryland.gov/support/Pages/open-data-portal.aspx

[6] https://opendata.maryland.gov/Demographic/Maryland-Counties-Socioeconomic-Characteristics/is7h-kp6x

[7] https://opendata.maryland.gov/Demographic/Choose-Maryland-Compare-Counties-Demographics/pa7d-u6hs

[8] https://opendata.maryland.gov/Housing/Choose-Maryland-Compare-Counties-Quality-Of-Life/dyym-bjv4

[9] https://opendata.maryland.gov/Education/Choose-Maryland-Compare-Counties-Education/63pe-mygy

[10] https://opendata.maryland.gov/Business-and-Economy/Choose-Maryland-Compare-Counties-Workforce/q7q7-usgm

[11] https://opendata.maryland.gov/Public-Safety/Violent-Crime-Property-Crime-by-County-1975-to-Pre/jwfa-fdxs

[12] https://en.wikipedia.org/wiki/List_of_counties_in_Maryland

[13] https://opendata.maryland.gov/Administrative/Maryland-Counties/g8er-va3s

[14] https://www.coursera.org/learn/applied-data-science-capstone/discussions/weeks/5/threads/OwNJTEq6TBiDSUxKunwYLw

[15] https://data.mongabay.com/igapo/zip_codes/MD.htm