Exploration of the Spread of COVID-19 by Socio-economic factors in Maryland

Eric Ohemeng

2023-12-13

Github: https://github.com/Erico250/Final-Project

INTRODUCTION

This project explores the dynamics between socioeconomic factors, and the spread of the COVID-19 pandemic across counties in Maryland. The research is guided by the question: "What patterns exist in COVID-19 cases across various socioeconomic and demographic characteristics within the State of Maryland in 2021?"

DATA COLLECTION

The dataset central to the project amalgamates several data sources, each contributing vital information to the analysis of COVID-19 across Maryland's counties. COVID-19 Data: Pulled from the Johns Hopkins University database available on Google BigQuery, this dataset provides granular, county-level details on the number of confirmed COVID-19 cases for 2021.
Socioeconomic and Demographic Data: Median household income and obtained from the American Community Survey (ACS) data, accessed via the Census API. Median age, also retrieved from the ACS, adds a demographic dimension to the project, facilitating an examination of how age distributions within counties might relate to pandemic patterns. Geographical Data: To visualize the geographic distribution of cases a, I use a shapefile outlining the boundaries of Maryland's counties. This GIS data underpins the creation of the choropleth map. SQL queries within the R environment were used to extract relevant COVID-19 data. The dplyr package in R facilitated data wrangling, including filtering, summarizing, and merging the datasets.

```
## # A tibble: 6 x 4
##   County         state    Cases deaths
##   <chr>          <chr>    <int>  <int>
## 1 Allegany       Maryland  6989    210
## 2 Anne Arundel   Maryland 43685    622
## 3 Baltimore      Maryland 65494   1537
## 4 Baltimore City Maryland 52759   1126
## 5 Calvert        Maryland  4209     78
## 6 Caroline       Maryland  2336     25
```

```
## # A tibble: 6 x 7
##   County    Cases deaths TotalPopulation MedianAge MedianHIncome PerCapitaIncome
##   <chr>     <int>  <int>           <dbl>     <dbl>         <dbl>           <dbl>
## 1 Allegany   6989    210           68684      41.4         51090           26762
## 2 Anne Aru~ 43685    622          584064      38.5        108048           51113
## 3 Baltimore 65494   1537          850702      39.4         81846           43290
```

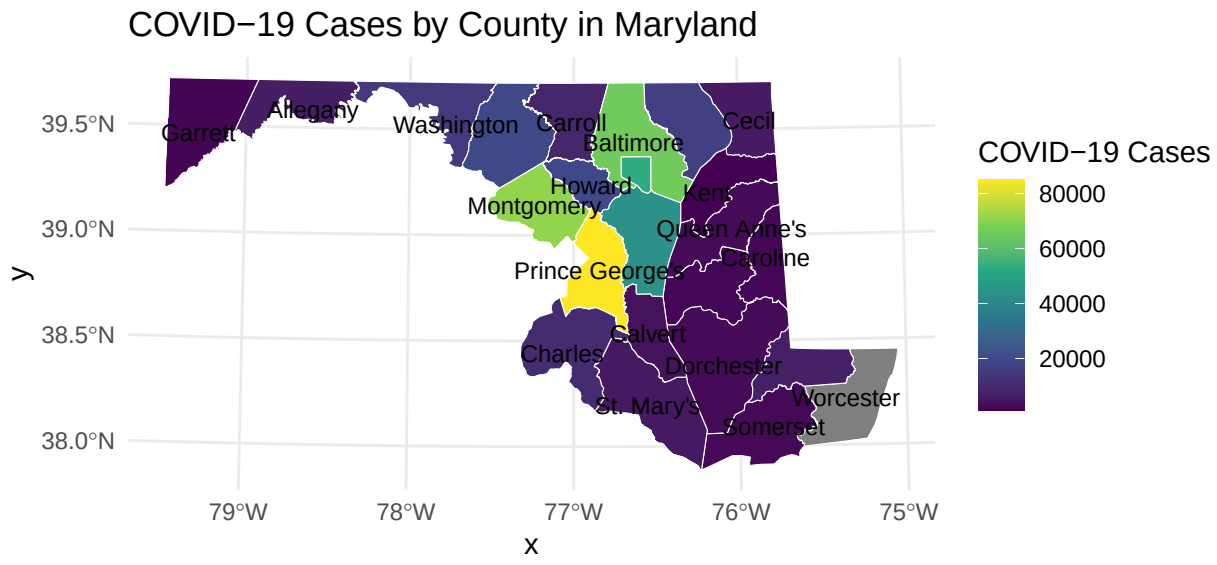
```
## 4 Calvert    4209     78          92515       40.2         120295          50496
## 5 Caroline   2336     25          33234       39.2          63027          32186
## 6 Carroll    9464    239         172148       41.7         104708          45800
```

ANALYSIS

The project incorporated a choropleth map to spatially analyze case distribution, a bar chart for cases rates per capita, and a scatter plot to evaluate the relationship between median household income, median age, and COVID-19, categorized by cluster analysis.

Choropleth Map

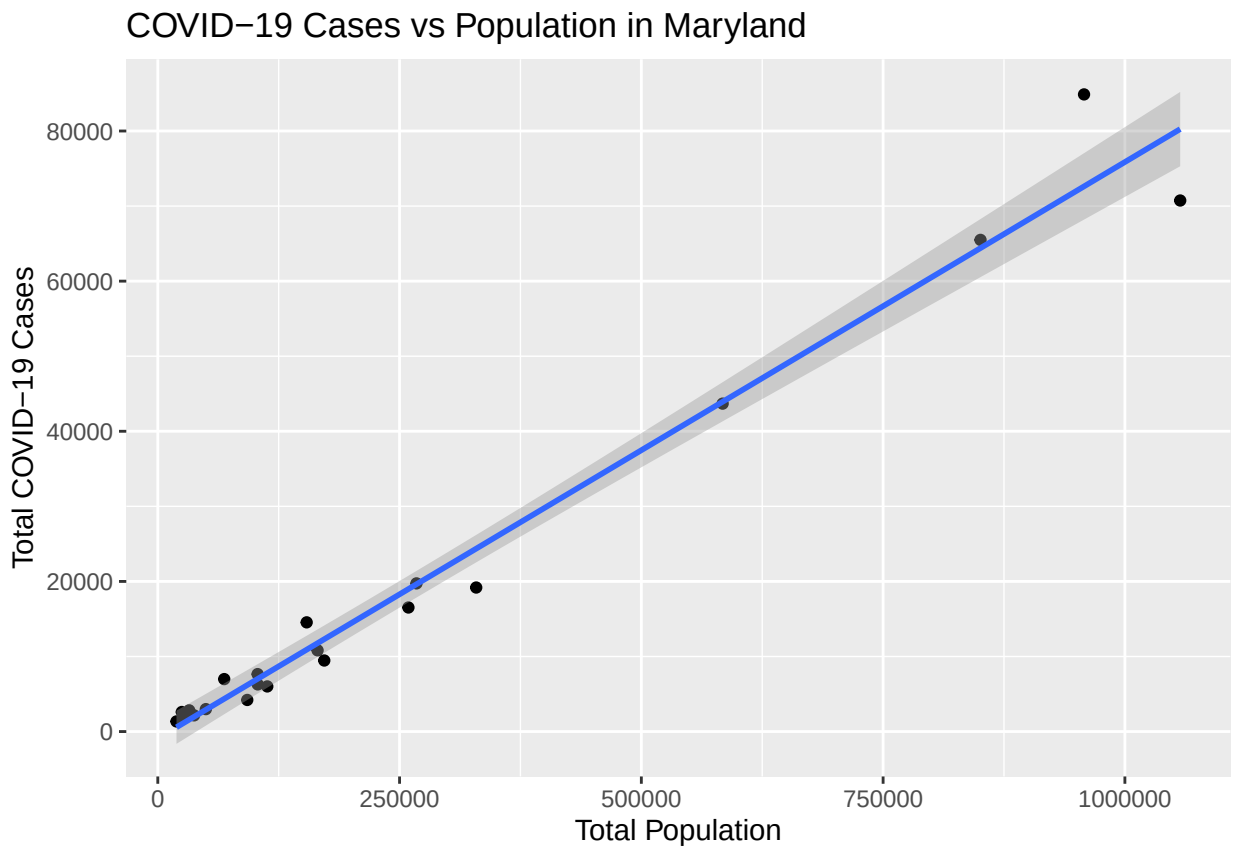
The map below shows the concentration of COVID-19 cases in urban counties such as Prince George's and Baltimore City, with darker shades indicating higher numbers. This suggests that densely populated areas experienced a more severe spread. Prince George's County stands out with its distinctive yellow shading, which, indicates the highest number of cases in the state. Rural Counties such as Somerset and Allegany County display lighter colors, indicating fewer cases. This could reflect several factors, including lower population density, potentially different levels of virus exposure



The scatter plot with a linear regression line, is showing a strong positive correlation between the total population and the number of COVID-19 cases in Maryland counties. This correlation coefficient suggests a very strong linear relationship; as the population in a county increases, the number of COVID-19 cases also increases. This visualization is helpful in understanding the relationship between population density and the spread of COVID-19, indicating that more populous areas may have had higher case numbers, which could be due to factors like greater person-to-person contact rates in more densely populated areas.

```
cor_pop <- cor(cleaned_data$Cases, cleaned_data$TotalPopulation)
cor_pop
```

```
## [1] 0.9874431
```

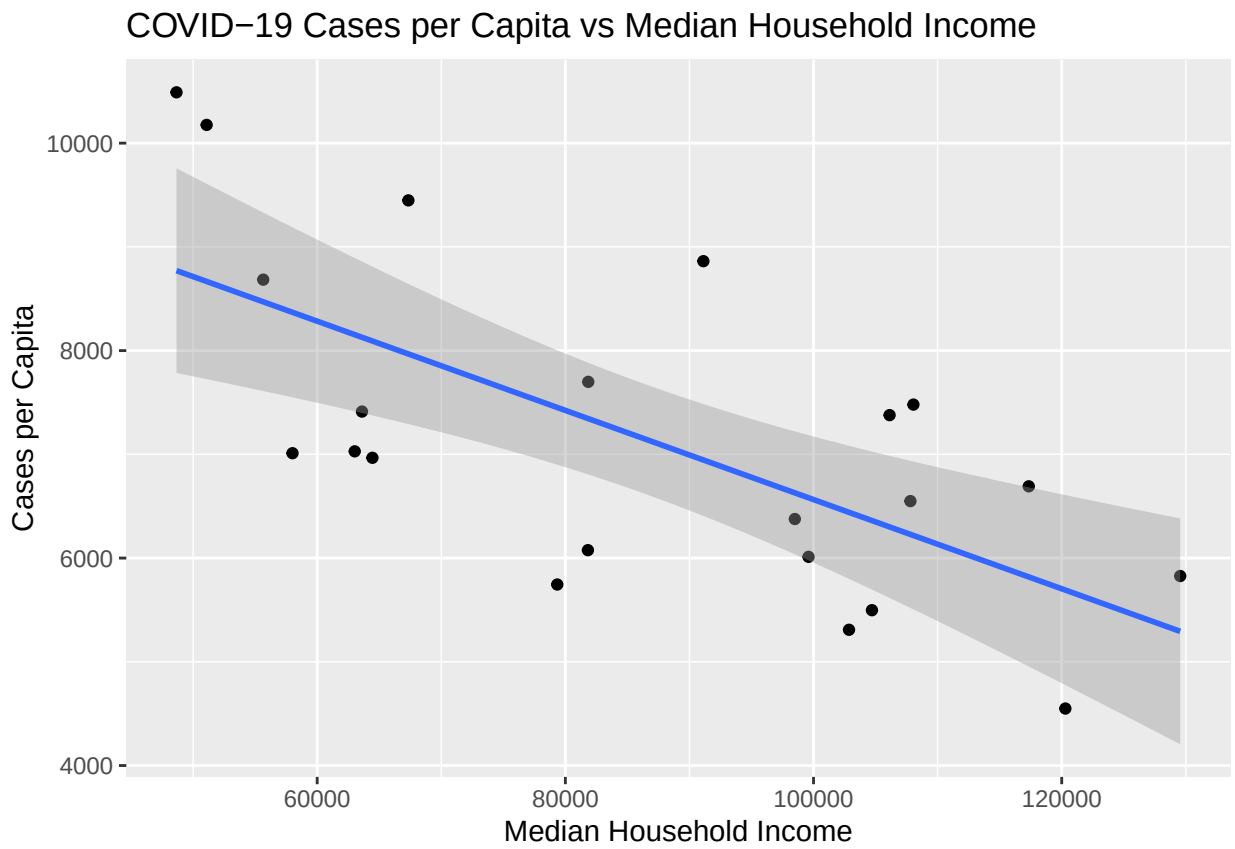


To be able to compare cases among counties, I computed per capita rates. The cases are multiplied by 100,000 to represent the standard rate per 100,000 inhabitants.

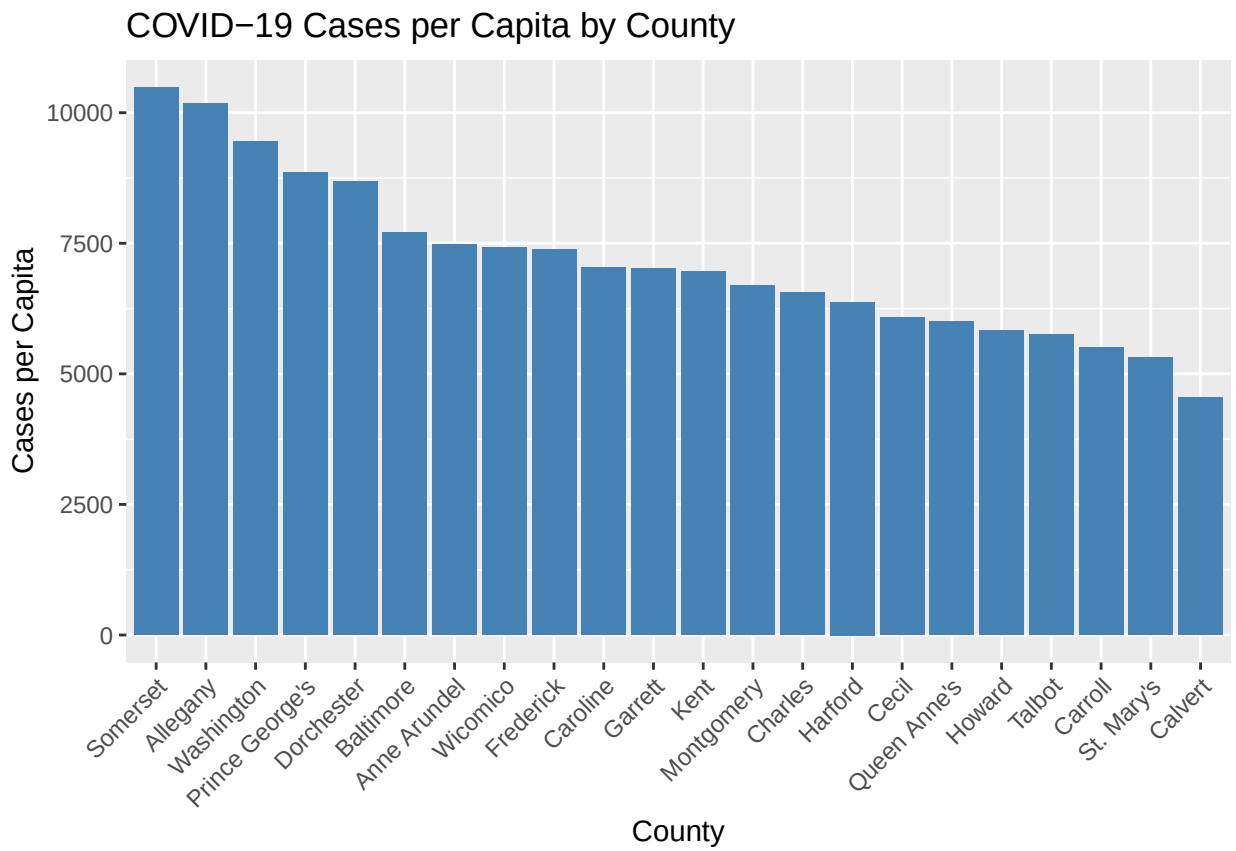
```
## # A tibble: 6 x 5
##   County       cases_per_capita deaths_per_capita MedianAge MedianHIncome
##   <chr>                   <dbl>             <dbl>     <dbl>         <dbl>
## 1 Allegany               10176.             306.       41.4         51090
## 2 Anne Arundel            7479.             106.       38.5        108048
## 3 Baltimore               7699.             181.       39.4         81846
## 4 Calvert                 4550.              84.3      40.2        120295
## 5 Caroline                7029.              75.2      39.2         63027
## 6 Carroll                 5498.             139.       41.7        104708
```

```
cor_results <- cor(covid_data$cases_per_capita, covid_data$MedianHIncome)
cor_results
```

```
## [1] -0.6726962
```



The bar chart below, representing standardized COVID-19 cases per capita, reveals that the per capita rate varies significantly between counties when adjusted for population. For instance, counties like Prince George that appeared to have a high number of total cases in the choropleth map above exhibits a moderate per capita rate due to a larger population base. Conversely, less populous counties such Allegamy with fewer total cases display a higher per capita rate, indicating a more substantial impact on the community. The standardization has unveiled a nuanced landscape of COVID-19's impact in Maryland, one that demands a tailored response that considers the unique socioeconomic and demographic fabric of each county.



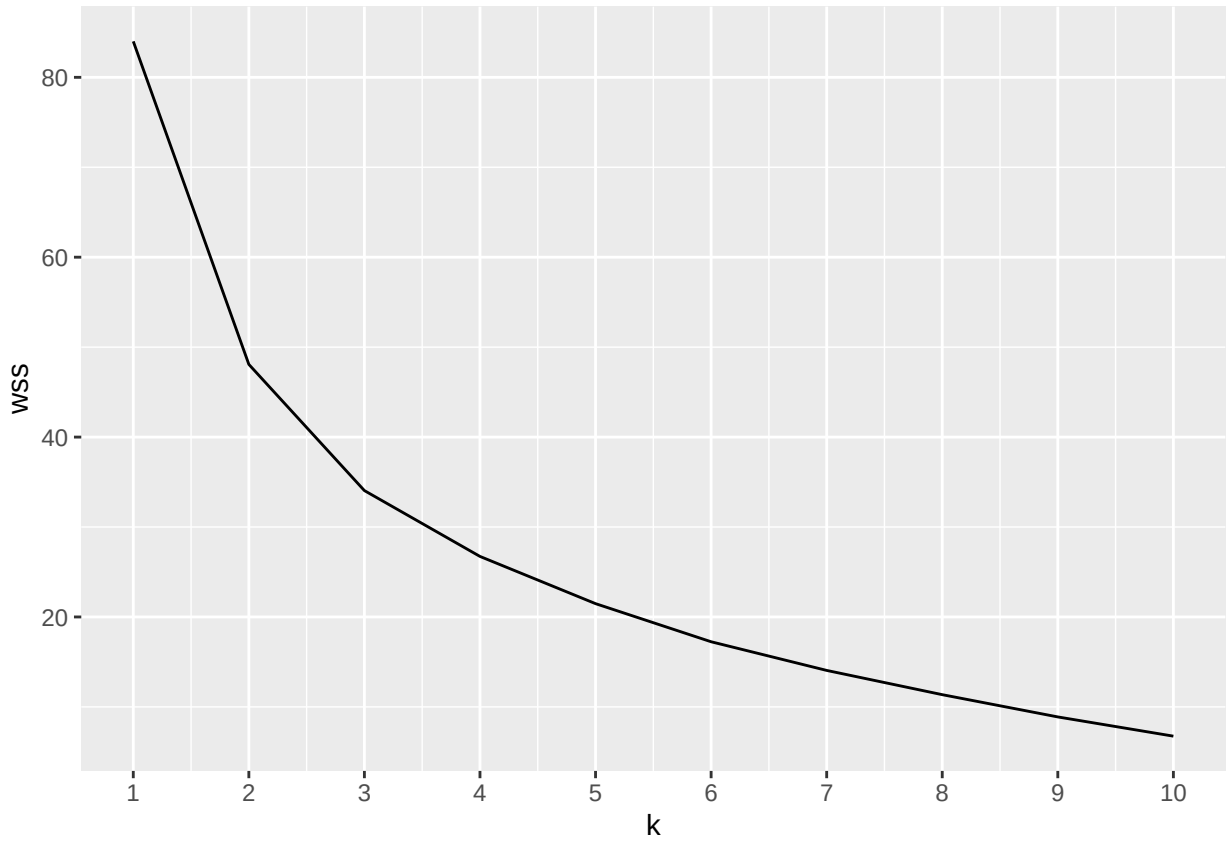
By employing clustering techniques, I aim to discern patterns in the infection rates per capita and understand how these may correlate with a county's economic standing and demographic profile. This approach allows me to explore the hypothesis that higher incomes and specific age demographics may impact the vulnerability of communities to the pandemic, providing insights that are crucial for informing public health strategies and interventions. Two scatter plots were generated from county-level data for COVID-19 cases per capita. One plot correlates with median household income and the other with median age. Each plot assigns counties to one of three clusters based on similarities in these attributes and their COVID-19 case rates.

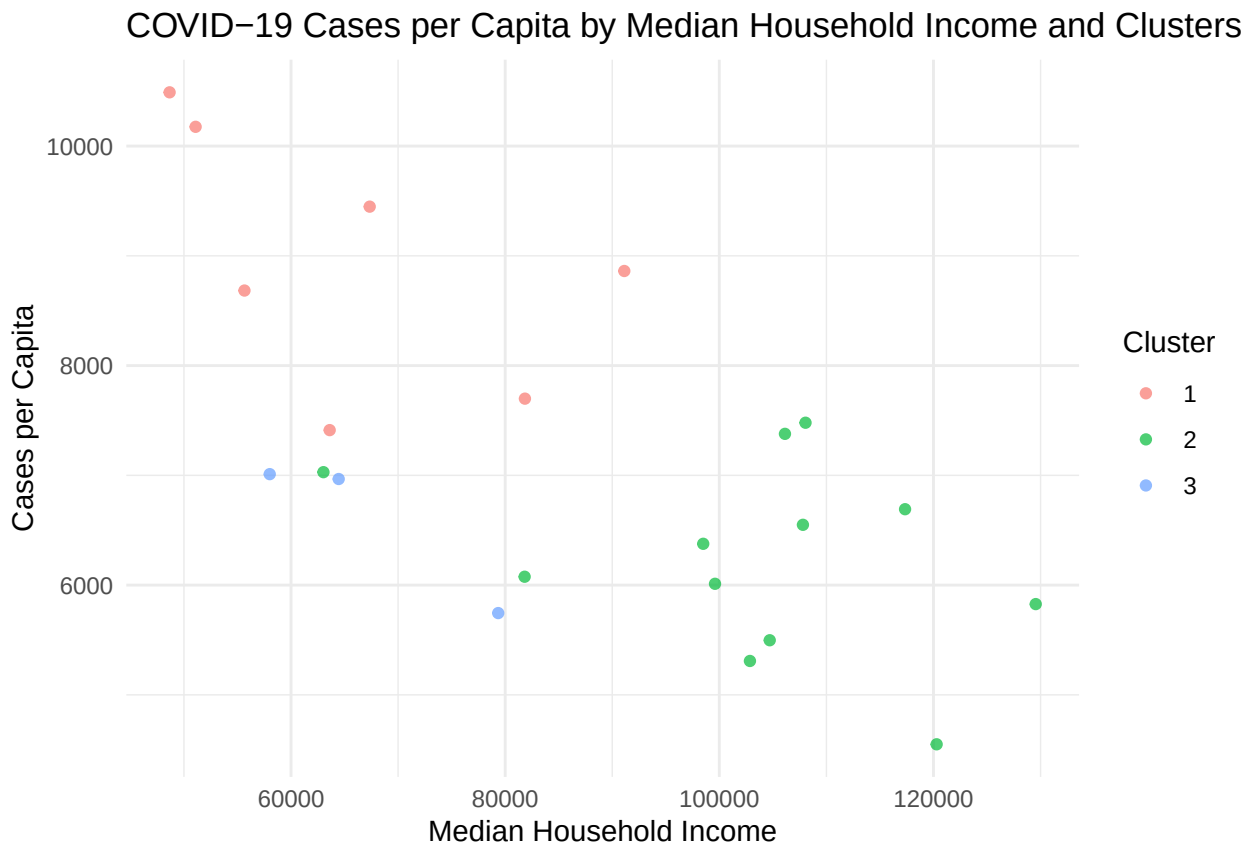
`Cluster Analysis by Median Household Income`

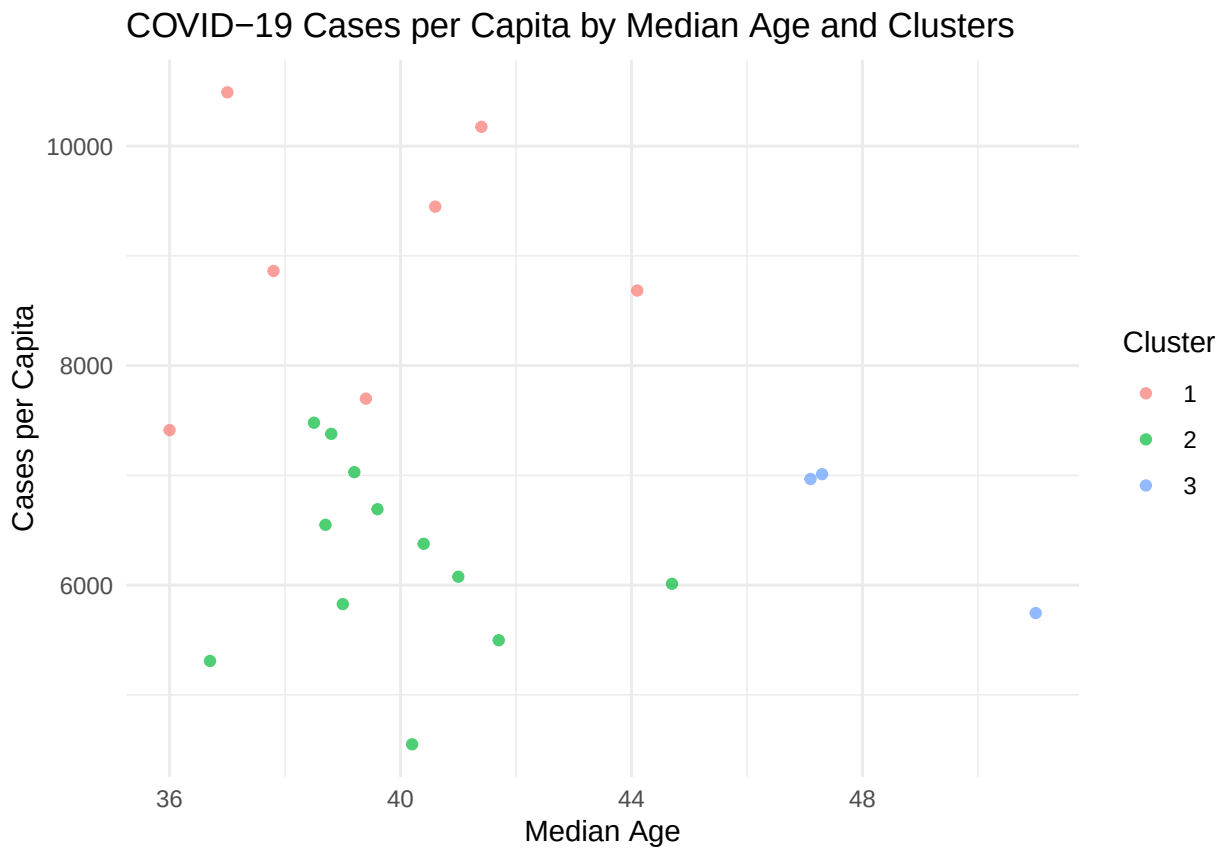
The scatter plot illustrates that counties in Cluster 1, marked by higher case rates per capita, are not strictly associated with lower household incomes, suggesting that income alone does not predict COVID-19 prevalence. Clusters 2 and 3, which consist of counties with moderate to lower case rates per capita, seem to include counties with higher median household incomes, hinting at a possible protective effect of economic status against the spread of COVID-19.

`Cluster Analysis by Median Age`

Similarly, the plot comparing median age with COVID-19 cases per capita shows that Cluster 1 encompasses a range of median ages, indicating that the median age by itself is not a decisive factor in COVID-19 case rates. Clusters 2 and 3 do not display a clear trend regarding median age, reinforcing the notion that other variables may also be influential.


80
60
40
20
wss
1
2
3
4
5
6
7
8
9
10
k


COVID−19 Cases per Capita by Median Household Income and Clusters
Cases per Capita
10000
8000
6000
60000
80000
100000
120000
Median Household Income
Cluster
1
2
3



CONCLUSION

The analysis of COVID-19 spread in Maryland, as depicted through bar charts, calculated correlation coefficients, and clustering by socioeconomic factors, reveals a nuanced picture of the pandemic's spread. The bar charts demonstrate a clear trend where counties with lower median household incomes recorded disproportionately higher COVID-19 cases per capita. The correlation analysis reinforces this observation, showing a negative correlation between income levels and COVID-19 cases per capita. Additionally, the clustering analysis divides counties into groups that suggest a potential protective effect of higher median household income against the spread of COVID-19. The median age, while included in the clustering, does not show as strong a correlation, indicating that income levels may be a more significant factor in understanding the distribution of COVID-19 cases across the state.

LIMITATIONS

The project, while revealing in its findings, is subject to several limitations. The correlational nature of the project precludes any assertion of causality between median income and median age and COVID-19 case rates. Furthermore, the exclusion of variables such as healthcare accessibility, and public health policies from the project could have significant, unaccounted effects on the observed patterns.Future studies would benefit from a longitudinal approach that captures these dynamics and employs statistical methods robust enough to infer causal relationships.