FACTORS INFLUENCING STATEWIDE COVID19 CASES

Research Questions

1. Do states with mask mandates tend to have lower numbers of COVID cases per 100k?
2. Do states with higher income levels tend to have lower COVID Case numbers per 100k?
3. Do states with a higher number of trips have increased numbers of COVID cases per 100k?
4. Do states with higher median age have more COVID Case numbers per 100k?

Background

According to the WHO, “Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.” The virus quickly became a pandemic with far-reaching effects. The death tolls have been massive and also highly variable. We decided to analyze several potential causes of higher cases and deaths for our project, including trips taken per day, income levels, the population of uninsured people, governor party affiliation, and whether or not the state has a mask mandate.

Data Collection

Independent Variables

Daily Trips: Quantitative data on The monthly average number of trips taken daily from March 2020 to December 2020. The data was collected from the Bureau of Transportation Statistics’ “Trips by Distance” dataset. We divided the total number of trips taken each day by the total population and averaged these numbers across our timeframe for each state.

Income: Quantitative data on median household income in each state from the US Census.

Health: Quantitative data on the percentage of uninsured residents in each state from the US Census.

Government Party: Qualitative data on Governor party affiliation from the National Governors’ Association, using “R” for Republican and “D” for Democrat.

Mask: Qualitative data on whether or not each state issued a mandatory order to wear masks from US News and AARP articles, using "Y" for Yes and "N" for No.

Stay at Home: Qualitative data on if a State had statewide lockdown measures, partial lockdown measures, or no lockdown measures from an NBC article, using "Y" for Statewide, "S" for Partial, and "N" for No.

Median Age: Quantitative data on median age in the United States in 2019, by state from Statista.

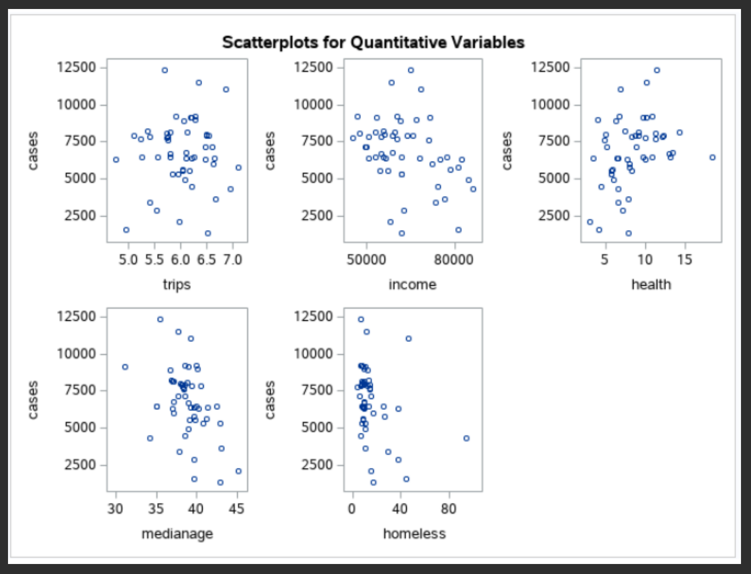
Homelessness: Quantitative data on the estimated rate of homelessness in the United States in 2019, by state from Statista.

Dependent Variable

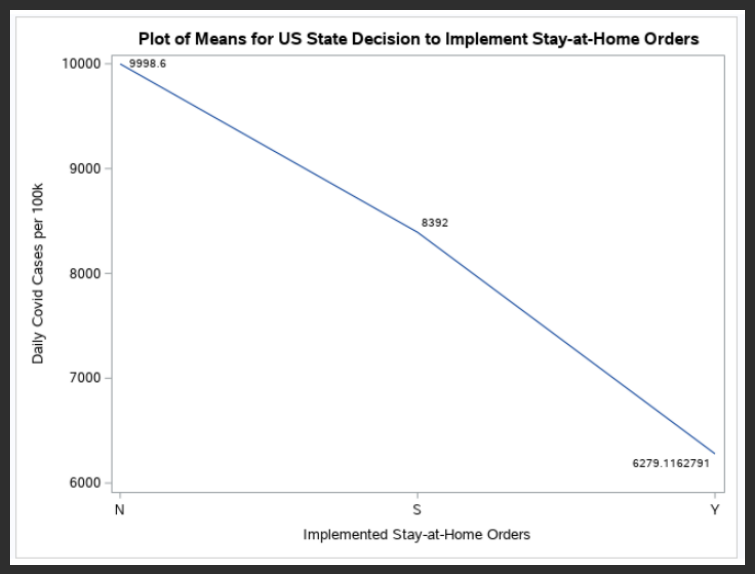
Cases: Used the CDC’s COVID tracking tool with the time period from January 21, 2020, to January 8, 2021, to determine the daily number of cases per 100,000 people in each state, and averaged the data.

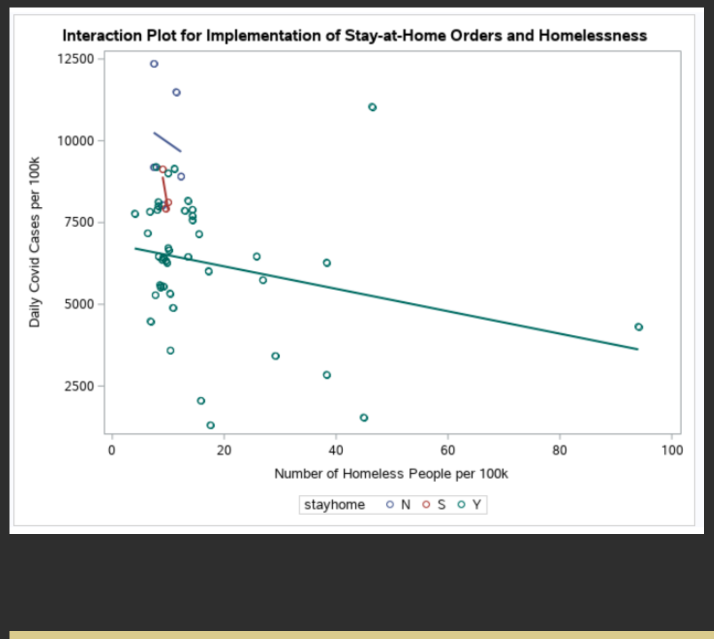
Exploratory Data Analysis

Scatterplots (Quantitative)



Plot of Mean (Qualitative)



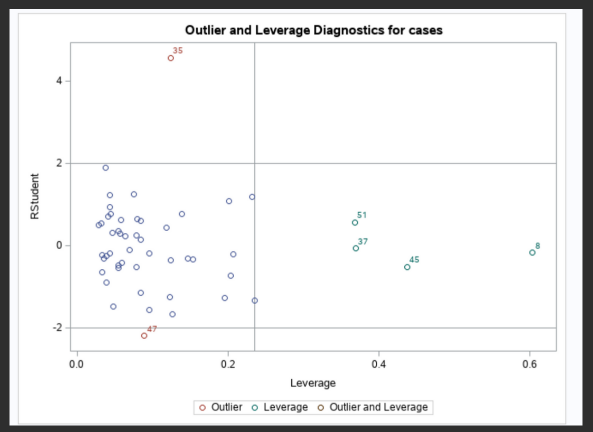


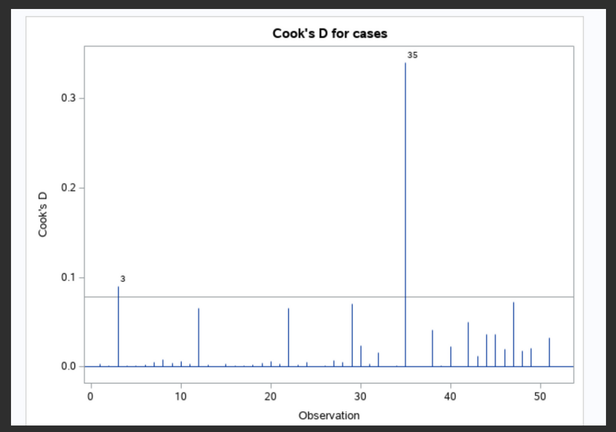
Interactions

The lines for the interaction between stay at home orders and homelessness have different slopes, therefore there likely is an interaction between the state's decision to implement stay at home orders and homelessness to influence the average daily cases per 100k.

ANALYSIS

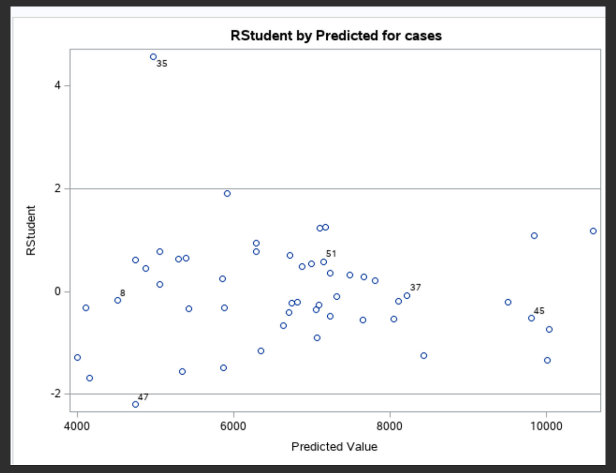
Model Assumptions: Our assumptions were not violated after analyzing the Mean Zero Assumptions, Constant Variance assumptions, Normality assumptions, and Independence assumptions.

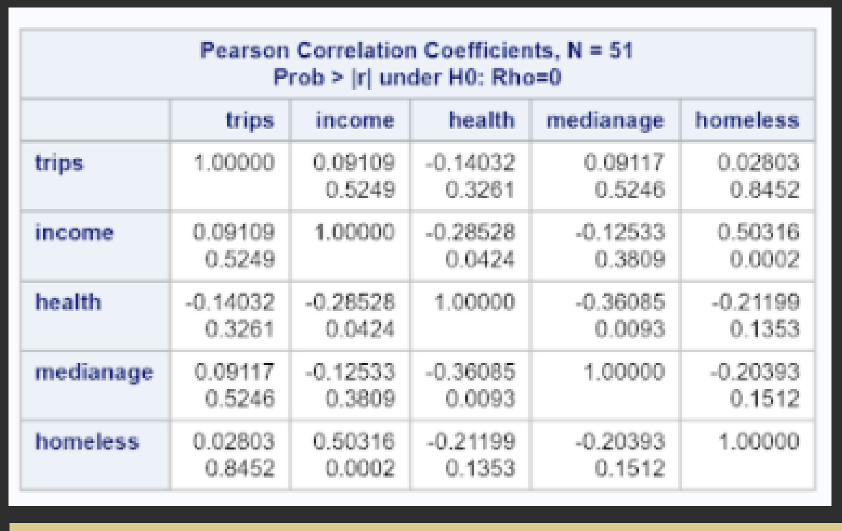

Outlier and Leverage Diagnostics for cases
RStudent
4
2
0
-2
35
45
47
0.0
0.2
0.4
0.6
Leverage
○ Outlier ○ Leverage ○ Outlier and Leverage



Observation 35 is influential as it is above the threshold for Cook’s distance, it is an outlier in the outlier/leverage graph, and its studentized residual is greater than 2.

Observation 47 is also influential as it is an outlier in the outlier/leverage graph and its studentized residual is less than -2.





Pearson Correlation Coefficients, N = 51
Prob > |r| under H0: Rho=0

	trips	income	health	medianage	homeless
trips	1.00000	0.09109 0.5249	-0.14032 0.3261	0.09117 0.5246	0.02803 0.8452
income	0.09109 0.5249	1.00000	-0.28528 0.0424	-0.12533 0.3809	0.50316 0.0002
health	-0.14032 0.3261	-0.28528 0.0424	1.00000	-0.36085 0.0093	-0.21199 0.1353
medianage	0.09117 0.5246	-0.12533 0.3809	-0.36085 0.0093	1.00000	-0.20393 0.1512
homeless	0.02803 0.8452	0.50316 0.0002	-0.21199 0.1353	-0.20393 0.1512	1.00000

Multicollinearity:

- Correlation and scatterplot matrixes do not indicate any significant linear relationships between any two variables.
- No VIF is above 10 and average VIF is less than 3.
- Multicollinearity will not be a problem.

Stage 1: Add Quantitative Predictors

Variable Screening: (stepwise selection SLEntry = 0.10 and SLStay = 0.10)
y = β0 + β1(income)+ β2(median age)+ β3(homeless)
Process: Variable screening showed that uninsured rate (variable 'health') and number of trips are not significant predictors. To further investigate this, we fitted a MLR model with all five quantitative predictors and performed individual t-tests for 'trips' and 'health', which confirmed the variable screening results.

Stage 2: Add Qualitative Predictors

Initial: y = β0 + β1(income)+ β2(median-age)+ β3(homeless) + β4(DummyParty) +β5(DummyMask) +β6(DummyHome1) +β7(DummyHome2) where DummyHome1 = {1 if "Y," 0 otherwise} and DummyHome2 = {1 if "S," 0 otherwise).
Process: Nested-F test and individual t-test showed mask mandate and governor party were not useful predictors, so they were removed from the model.
New: y = β0 + β1(income)+ β2(median-age)+ β3(homeless) + β6(DummyHome1) +β7(DummyHome2)

Stage 3: Add Interactions

Initial: y = β0 + β1(income)+ β2(median age)+ β3(homeless) + β6(DummyHome1) +β7(DummyHome2) + β8(DummyHome1)*(homeless) + β9(DummyHome2)*(homeless) where DummyHome1 = {1 if "Y," 0 otherwise} and DummyHome2 = {1 if "S," 0 otherwise).
Process: Based on EDA plots, we added the interaction we though was strongest into the model, which was between homeless rate and the stay at home order. However, the interaction was insignificant based on the nested F-test, so we removed this interaction from the model.
FINAL MODEL: y = β0 + β1(income) + β2(median-age) + β3(homeless) + β6(DummyHome1) +β7(DummyHome2)

Least Squares Prediction Equation

ŷ = 29987 - 0.0537*(income) - 440.871*(median age) - 34.734*(homeless) - 2442.518*(DummyHome1) - 2268.324*(DummyHome2)

where homedum1 = {1 if "Y," 0 otherwise} and homedum2 = {1 if "S," 0 otherwise).

- **0.0534** daily cases per 100k decrease for every increase in median household income by one dollar.
- **440.871** daily cases per 100k decrease for every increase in median age by one year.
- **34.734** daily cases per 100k decrease for every one person increase in homeless people per 100k.
- **2442.518** daily cases per 100k decrease if US state had statewide stay-at-home orders.
- **2268.324** daily cases per 100k decrease if US state had partial stay-at-home orders.

Conclusion

- Median household income, median age, homelessness, and type of stay-at-home order implemented were significant predictors.
- The following were not significant predictors: trips taken, uninsured residents, political party, and mask mandates.
- Model is overall decent for predicting average daily COVID-19 cases per 100k for a US state but could use some improvement in more exploration of certain parameters.

Data/Analysis Limitations

- Always the possibility that the sample populations are not representative of the true population despite coming from reputable governmental databases.
- Data is not always from the same time period, with some of our variables coming from years prior to the COVID 19 Pandemic
- Difficulty determining which predictors were significant as the Nested F test showed income along with 4 others as not significant but our stepwise variable screening process opted to include income in our analysis.

Future Research

- Consider other explanatory variables such as percent of state population with respiratory diseases or diabetes while using COVID-19 deaths as a new response variable for comparison.
- Make a logistic regression model using data of a random sample of individuals with COVID-19 with the response variable being whether or not they died from the virus.

Works Cited

- Source 0: WHO on Covid:
  - https://www.who.int/health-topics/coronavirus#tab=tab_1
- Source 1: Trips: (Website Filter Condition: When Level is State)
  - https://data.bts.gov/Research-and-Statistics/Trips-by-Distance-2020/dac6-p3ut/data
- Source 2: Income: (The Table Data available under: By State)
  - https://www.census.gov/search-results.html?q=average+per+capita+income&search.x=0&search.y=0&search=submit&page=1&stateGeo=none&searchtype=web&cssp=SERP
- Source 3: Health: (Page 20, The Column of Data available under: 2019 Uninsured)
  - https://www.census.gov/search-results.html q=average+per+capita+income&search.x=0&search.y=0&search=submit&page=1&stateGeo=none&searchtype=web&cssp=SERP
- Source 4: Party:
  - https://www.nga.org/governors/
- Source 5: Masks:
  - http://www.usnews.com/news/best-states/articles/these-are-the-states-with-mask-mandates
  - https://www.aarp.org/health/healthy-living/info-2020/states-mask-mandates-coronavirus.html
- Source 6: Cases:
  - (Data available under “Data Table for Case Rate by State/Territory” -- Website Conditions: View = Cases; Data Table for Case Rate by State/Territory and Metric = Rate per 100,000)) https://data.bts.gov/Research-and-Statistics/Trips-by-Distance-2020/dac6-p3ut/data