# When Will My Reflection Show Who I Am Inside?

Social Distancing Measures as a Reflection of Political Party Preference

Gene Ho, Hannah Qi, Jasmine Wu

## Abstract

This project uses different methods and experiments to explore the relationship between the implementation of social distancing measures and democratic to republican ratio for each county in the U.S. The data was obtained from Johns Hopkins University. COVID-19 research is being rapidly published, as it is an important public health crisis. Linear and logistic regressions were performed to find models for this between political preference and social distancing order implementation. The final logistic and log-linear models performed fairly well on our test set. Assuming the coefficients for these models are significant, they should be suitable for modeling the outcomes' relationship to the selected features.

## Introduction

On January 30, 2020, the World Health Organization (WHO) issued a "public health emergency of international concern," in response to the discovery of a novel coronavirus (COVID-19) in Wuhan, China. Since then, the virus has rapidly spread across the globe, with 4,088,848 confirmed cases and 283,153 deaths across 212 countries and territories. COVID-19 belongs to the same family as two other influenza-like viruses that caused pandemics in the early 2000s: SARS (Sudden Acute Respiratory Syndrome) and MERS (Middle Eastern Respiratory Syndrome). Within a month of the initial infection, COVID-19 surpassed the final count of SARS cases in 2003.

Due to the rapid spread of this virus all over the world, everyone has a role to play to reduce and slow the transmission. Social distancing is an essential step in preventing the spread of COVID-19. The predominant political parties have different perspectives on how best to handle this pandemic: "In Florida, Georgia, Texas, and Missouri, for example, local officials in big Democratic-leaning cities (such as Miami, Atlanta, Dallas, and St. Louis) have moved to severely restrict activity. But those mayors have complained that their efforts are being undercut by the refusal of Republican governors—whose support is typically greatest in rural areas—to impose uniform limits across the state" according to Ronald Brownstein.

## Research Question and Hypothesis

Research Questions:

How does the timing of the implementation of various social distancing measures (i.e. time to issue) predict a county's democrat to republican ratio?

How does the timing of the implementation of various social distancing measures (i.e. time to issue) predict a democratic/republican majority in a county?

Hypothesis:

Counties that took longer to implement social distancing measures since the dates of their first case have lower Democrat to Republican ratios.

## Methods and Experiments

### Data Cleaning

We utilized two separate datasets for our analysis: `abridged_counties.csv` and `time_series_covid19_confirmed_US.csv`.

The research question involved time and dates, so the columns in `abridged_counties.csv` corresponding to the order issue date (i.e. `'stay at home'`, `'>50 gatherings'`, `'>500 gatherings'`, `'public schools'`, `'restaurant dine-in'`, `'entertainment/gym'`) were converted from proleptic Gregorian ordinal to a `pandas Timestamp`. Some of the counties were missing state abbreviations, so a dictionary from Github was utilized to map state names to state abbreviations.

Then, the two dataframes were joined on FIPS, which is a unique identifier for each county. After joining, only the relevant columns were kept. Counties that did not have a reported value for the outcome variable (`‘dem_to_rep_ratio’`) were removed from the dataset. The date of the first case for each county was identified by defining the `first_case()` function and also converted them to a `pandas Timestamp`. The data in both datasets included information only up until 4/18/20, so missing dates were imputed to be 4/19/20 to reflect it being outside of the scope of the data without introducing excessive bias. Table 1 describes the columns of `final_df`.

The predictor variables were the number of days in relation to the day of the first reported case of COVID-19. Since all dates were already in `pandas Timestamp` format, the order issue date was subtracted from the first case date to get the time to issue for each order. Negative values indicate that a county was proactive, instead of reactive, in these prevention measures. `pandas Timedelta` class appended “days” to the time difference; this was removed by converting it to a `datetime` class.

### Exploratory Data Analysis

Our exploratory data analysis was distinguished based on if it was on the predictor or outcome variable(s).

A grouped bar chart was created to map the “order of orders,” or the sequence of issuance for each social/physical distancing measure by each county (Fig 1). The stay at home order was by far the most common first measure implemented by US counties. The closure of public school was often the last order that was issued. There were virtually no stay at home and gym/entertainment venue closure orders as the last order.

`‘dem_to_rep_ratio’` was plotted as a histogram and box plot (Fig 2). The histogram was extremely right skewed, and extreme outliers were not readily apparent. However, the box plot showed that there were plenty of outliers, with a maximum of 22.23 democratic voters for every Republican voter. The minimum value was 0.03, which is 741 times smaller than the maximum--a very large range. There were a total of 265 counties that were considered outliers.

Two choropleths were created involving our outcome variable (`‘dem_to_rep_ratio’`) and its transformations. The first (Fig 3) was the log transformed ratio for each county in the United States. The dichotomized ratio was also plotted to show which political party had a majority in each county. In the first choropleth, there aren’t any regions that are extremely polarized towards one party or the other. However, there is a trend of the east and west coasts being more blue and this transitions to red closer to the interior of the country.

When looking at the second choropleth (Fig 4), there are many more counties that have a Republican majority than Democratic majority. The previous observations of Democratic counties being along the coasts still holds. This choropleth was not directly weighted by population size. However, when comparing it to a third choropleth (Fig 5) with county populations, more populous counties tended to have a Democratic majority.

Two correlation matrices were created. The first was with all six original covariates along with the outcome and its logarithmic transformation (Fig 6). There was relatively low correlation between the covariates and the two potential outcomes and relatively high correlation between each of the covariates. This caused some concerns regarding multicollinearity.

## Model Building

A training and test set were generated with the county-level data. Cross-validation was not necessary since it is typically used to maximize the “value” of each data point in a small dataset. However, we had a relatively large dataset of 3,114 counties/data points. Utilizing a 90:10 training:test split would still leave 311 counties to measure our model’s performance. While it was not necessary, we respecified the code for the split for each model to update the variable names, but the random seed was set so the individual counties in each set were the same.

### Linear Regression

The first model (`‘linear_model’`) was created with the following columns as the features of the X, `'stay at home', '>50 gatherings', '>500 gatherings', 'public schools', 'restaurant dine-in', 'entertainment/gym'`. `‘dem_to_rep_ratio’` is the Y.

A more selective model (`‘select_linear_model’`) was made by calculating variance inflation factors (VIF), which served as a measure of multicollinearity. Through this measure, the following covariates were identified to introduce the least collinearity: `'stay at home', '>500 gatherings', 'public schools', 'entertainment/gym'`. Fig 7 depicts the updated correlation matrix using only these four covariates and the two possible outcome variables. `‘dem_to_rep_ratio’` column was the outcome variable for this model as well.

A log-linear model (`‘log_linear_model’`) using the same features as `‘select_linear_model’`, but `‘dem_to_rep_ratio’` (outcome) was log-transformed in order to make our outcome have a more normal distribution and to account for the large magnitude of difference. When the histogram and boxplot were recreated for this log transformation (Fig 2), the histogram appeared more normal, and outliers were distributed on both ends of the boxplot whiskers.

### Logistic Regression

The `majority` column was created to answer the second research question of being able to use the same columns to predict a democratic or republican majority, a slightly more general question that can be answered with a logistic regression since it is a dichotomous outcome. It was created so that it would equal 1 if there was a republican majority (i.e. `‘dem_to_rep_ratio’` > 1) or 0 if there was a democratic majority (i.e. `‘dem_to_rep_ratio’` < 1).

**Results**

Error and Accuracy Metrics

The training error/accuracy for the three models are reported in Table 1. The log-linear model had a much lower error than the other linear model. For the logistic models, there were minimal gains in accuracy through the exclusion of the two variables. To avoid overfitting the data and thus improve generalizability, the linear and logistic regression models with only four of the six original covariates were considered to be our final models.

Log-Linear Model

$$\begin{aligned}\log(\texttt{dem\_to\_rep\_ratio}) = &-0.328 - 0.016 * \texttt{stay at home}\\ &+ 0.012 * \texttt{>500 gatherings} + 0.032 * \texttt{public schools}\\ &- 0.001 * \texttt{entertainment/gym}\end{aligned}$$

This model predicted that, for every additional day delay in:

- issuance of a stay at home order, there was an additional 0.98 ( $e^{-0.016}$ ) Democratic voter for every Republican voter...
- issuance of a ban on gatherings over 500 people, there was an additional 1.01 ( $e^{0.012}$ ) Democratic voter for every Republican voter...
- public school closures, there was an additional 1.03 ( $e^{0.032}$ ) Democratic voter for every Republican voter...
- gym/entertainment venue closures, there was an additional 0.99 ( $e^{-0.001}$ ) Democratic voter for every Republican voter...

...holding other covariates constant.

Logistic Model

$$\begin{aligned}\operatorname{logit}(\texttt{majority}) = &0.695 + 0.049 * \texttt{stay at home}\\ &- 0.013 * \texttt{>500 gatherings} - 0.089 * \texttt{public schools}\\ &- 0.025 * \texttt{entertainment/gym}\end{aligned}$$

This model predicted for every additional day in delay

- in issuance of a stay at home order, there was a 0.049 increase in the log odds, or 1.05 times the odds, that a county was Republican...
- in issuance of a ban on gatherings over 500 people, there was a 0.013 decrease in the log odds, or 0.98 times the odds that a county was Republican...
- in public school closure, there was a 0.089 increase in the log odds, or 0.91 times the odds that a county was Republican ...
- in issuance of a stay at home order, there was a 0.025 decrease in the log odds, or 0.98 times the odds that a county was Republican ...

...holding the other covariates constant.

Comparing the Best Model to the Testing Set

When the models were applied to our test set, the models performed better than on the training sets. The log-linear model had an error of 0.666 (compared to 0.68, lower is better), while the logistic regression model had an accuracy of 0.865 (compared to 0.85, higher is better).

### Discussion

There are some limitations with the data provided. For example, we only have the data for the estimated total population of the county in 2018 (Fig 5) which isn't representative data for the population during COVID-19. However, `PopulationEstimate2018` can give us a rough idea of the distribution of population in the United States.

None of the features were particularly interesting because they all had extremely low coefficients, making the significance of each somewhat questionable. However, the log transformation of the `‘dem_to_rep_ratio’` column was more interesting since it resulted in a model with a lower error than the other models. Overall, the attribute of measuring the time difference between first case and order issuance was interesting and avoided the need for a time-to-event analysis. As discussed earlier, `'>50 gatherings'` and `‘restaurant dine-in’` were excluded due to high VIFs, implying their introduction of collinearity.

A challenge we had with our data was handling `pandas Timestamp` objects. `pandas Timestamp` and `datetime` objects are not interoperable, so we had to make sure all our dates were of the same type. Another challenge we had was making sure we got the date of the first case for every county without any missing. We had to take into account counties that did not have a recorded date of the first case, so we gave these counties the date 4/19/20, the day after the last date in our data. This introduces some bias because we do not know the actual date of when that order was implemented if implemented at all. We also assumed that if a county did not have a record for the date of the first case, we gave those counties 4/19/20 as their date of the first case. Typically, imputed values are much greater than the range, but due to lack of content expertise, we wanted to limit the amount of error intentionally introduced.

Another bias we have in this project is that the data provided from the `abridged_counties` table only gives us the ratio of the number of votes received by the Democratic candidate over that received by the Republican candidate in 2016, which doesn’t include other parties and the data is not up to date. If we have more current data, our model and prediction will be more accurate. In addition, even though the U.S. is dominated by the Democratic Party and the Republican Party, it is necessary to address all the people in the country. A potential ethical dilemma is due to an inclusion of an individual in the confirmed and death counts due to COVID-19. They likely did not consent to being a part of this dataset, but since this is a reportable communicable disease during a public health emergency, they are automatically included. Although this data is aggregated on a county-level, alleviating most ethical concerns about identifiable data, there is a concern in how this data is being collected en masse.

The dates we have from the data are from 1/22/2020 to 4/18/2020, which is not up to date, Our data is restricted to only include information between 1/22/2020 and 4/18/2020, restricting our ability to model as new data is being produced and released daily. If more recent data was provided, we would have more data for the testing and training sets, which could help us predict a better model. It would also be interesting to have Trump approval ratings by county to see if there is a correlation with social distancing order implementation/

An ethical concern is that we do not know how Johns Hopkins University obtained the data or if anything else happened to the data from when Johns Hopkins University created the data to when it reached us. Also, our data--albeit weakly and assuming reverse causation--suggests Republican counties are slightly more likely to delay the implementation of social distancing orders. If this further implies they will have more cases, should these counties be eligible for increased testing capacity and funding even though they delayed preventative measures?

## Conclusion

Based on our results, we were able to derive a formula that relates the timing of implementation of social distancing orders to county-level political party preference. These coefficients of the model are not in the same direction (positive or negative) for all of the orders, suggesting there might be more to be considered when attempting to find an association between these variables. We did not conduct any significance tests on our coefficients, so we cannot be certain that our coefficients are actually statistically significant. This is especially a concern for us because the coefficients are so small. This is potentially something that can be done in future analyses since statistical significance tests are outside the scope of this class and are not natively supported by scikit-learn.

This project has prompted several other questions related to political preference and handling of COVID-19. One such question that we would be interested in investigating relates to how a county's handling of COVID-19, not necessarily limited to implementation of social distancing orders, relates to President Trump's approval rating.

## Presentation

Google Slides

Video Presentation

## Tables and Figures

### Tables

Table 1. Columns in `final_df` dataframe

| | |
|---|---|
| `county_FIPS` | state-county FIPS Code |
| `CountyName` | county name |
| `StateName` | state name |
| `dem_to_rep_ratio` | ratio of the number of votes received by the Democratic candidate over that received by the Republican candidate in the 2016 presidential election |
| `log_ratio` | natural logarithm of dem_to_rep_ratio |
| `majority` | political party majority based on dem_to_rep_ratio, such that Democratic majority = 0 and Republican majority = 1. |
| `Date of First Case` | date of first reported COVID-19 case in county |
| `stay at home` | number of dates between first reported COVID-19 case in county and implementation of stay at home order (negative values indicating proactive issue) |
| `> 50 gatherings` | number of dates between first reported COVID-19 case in county and ban on gatherings over 50 people (negative values indicating proactive issue) |
| `> 500 gatherings` | number of dates between first reported COVID-19 case in county and ban on gatherings over 500 people (negative values indicating proactive issue) |
| `public schools` | number of dates between first reported COVID-19 case in county and closure of public schools (negative values indicating proactive issue) |
| `restaurant dine-in` | number of dates between first reported COVID-19 case in county and ban on dining in at restaurants (negative values indicating proactive issue) |
| `entertainment/gym` | number of dates between first reported COVID-19 case in county and closure of entertainment venues and gyms (negative values indicating proactive issue) |

Table 2. Model Performance Metrics

| Model | Training Error (Linear) / Accuracy (Logistic) |
|---|---|
| Linear with 6 features | 0.818 |
| Linear with 4 selected features | 0.823 |
| **Log-Linear with 4 selected features** | **0.689** |
| Logistic with 6 features | 0.850 |
| **Logistic with 4 selected features** | **0.856** |

Figures

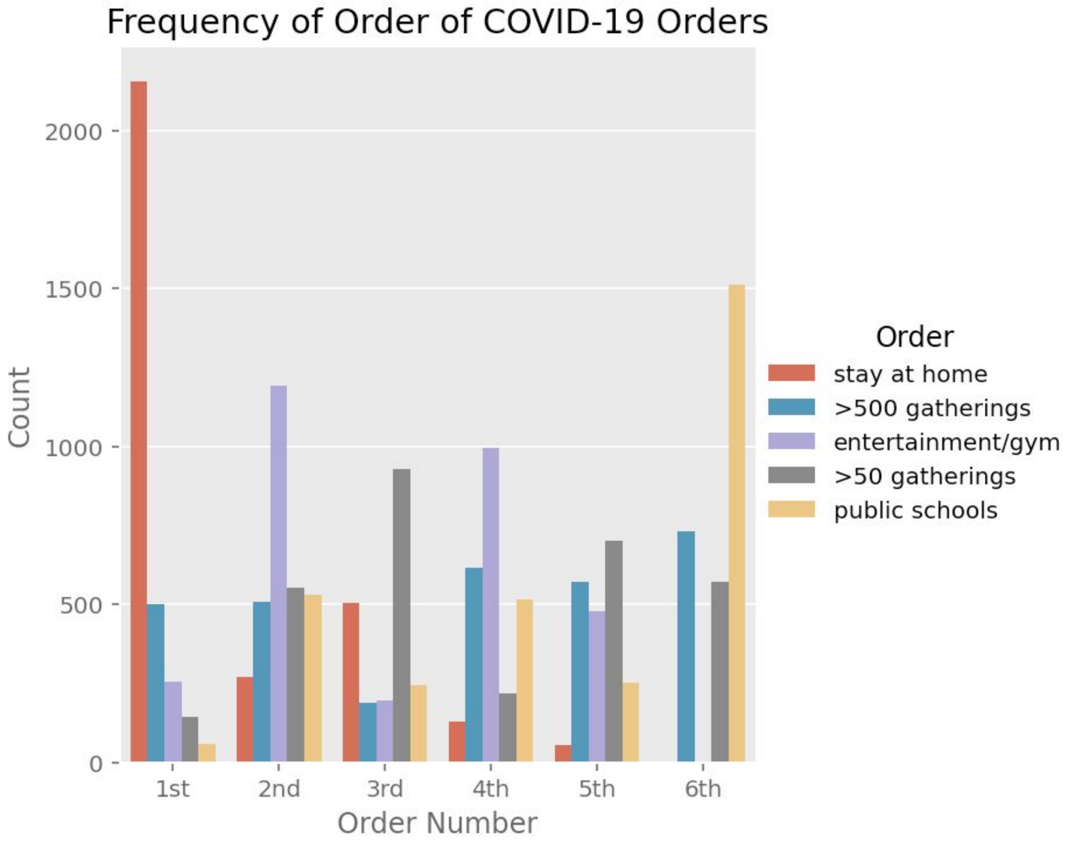

Fig 1. Order of Orders

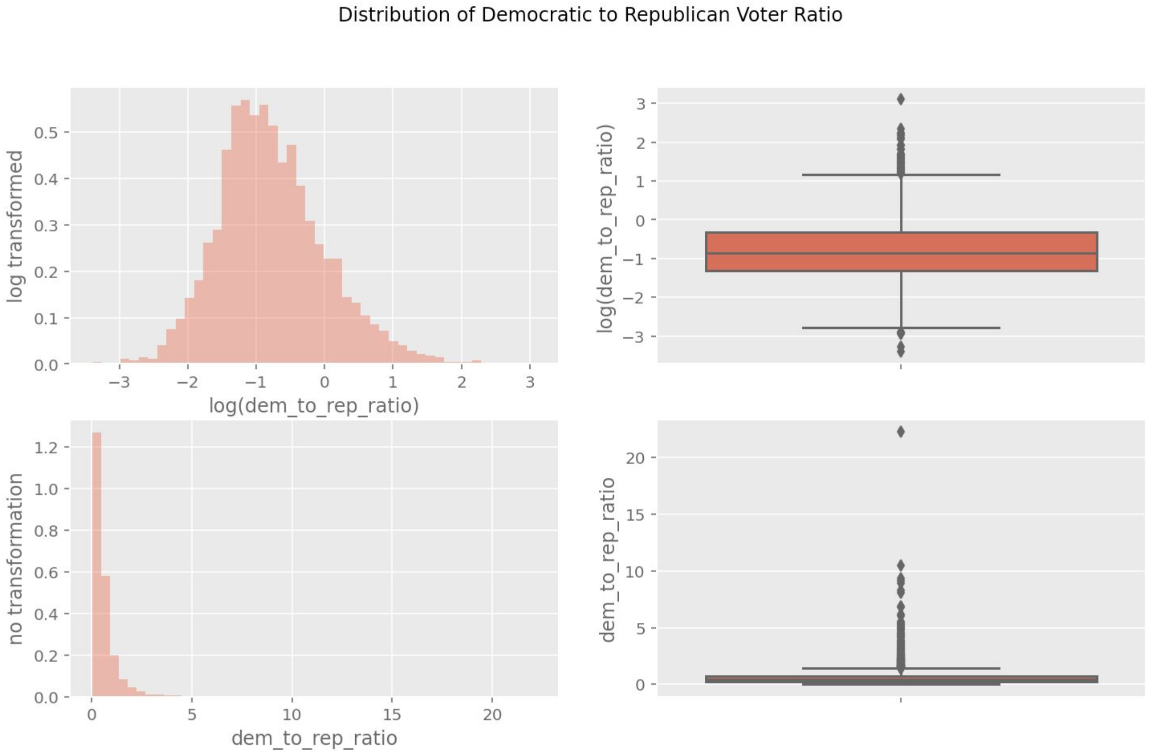

Fig 2. Distribution of Democrat to Republican voter ratio by US county

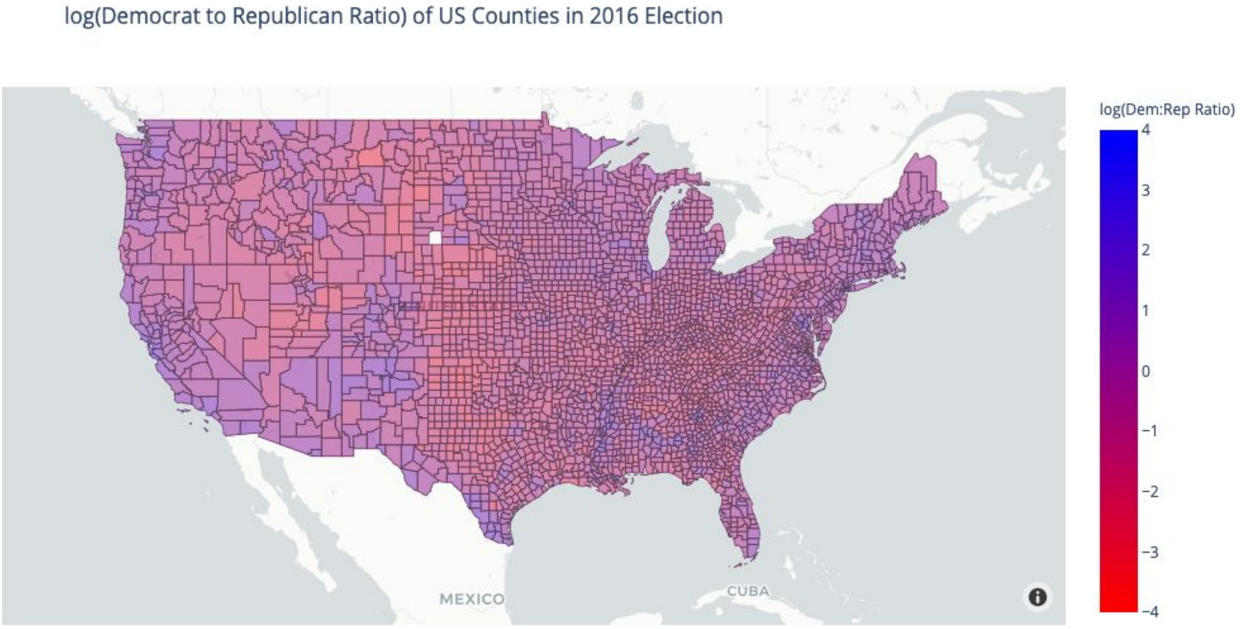
log(Democrat to Republican Ratio) of US Counties in 2016 Election
log(Dem:Rep Ratio)
4
3
2
1
0
−1
−2
−3
−4
MEXICO
CUBA

Fig 3. Map of Democrat to Republican voter logarithmic ratio by US county

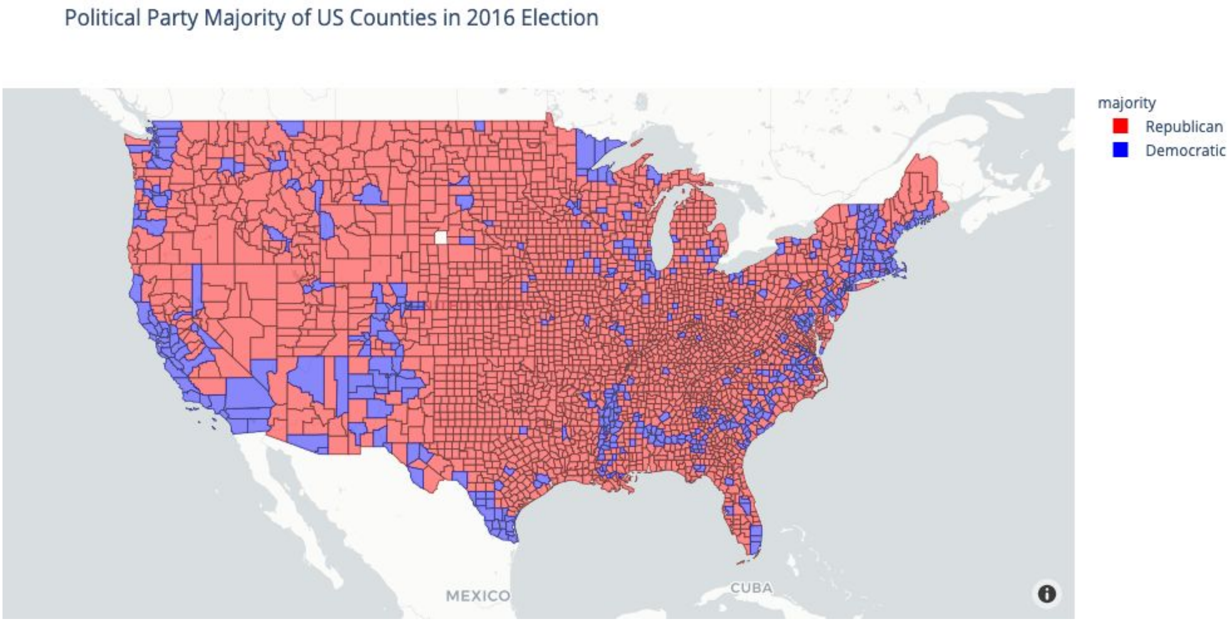

Fig 4. Map of Political Party Majority, based on number of votes in the 2016 Presidential election, by US County

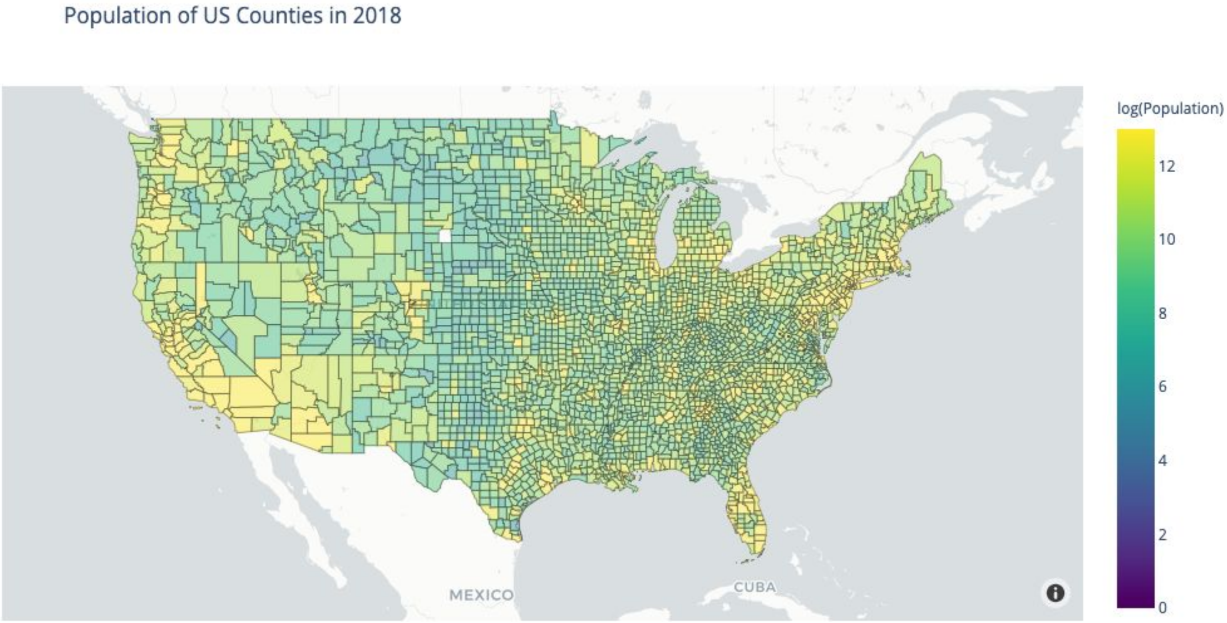

Fig 5. US county population, transformed by natural logarithm

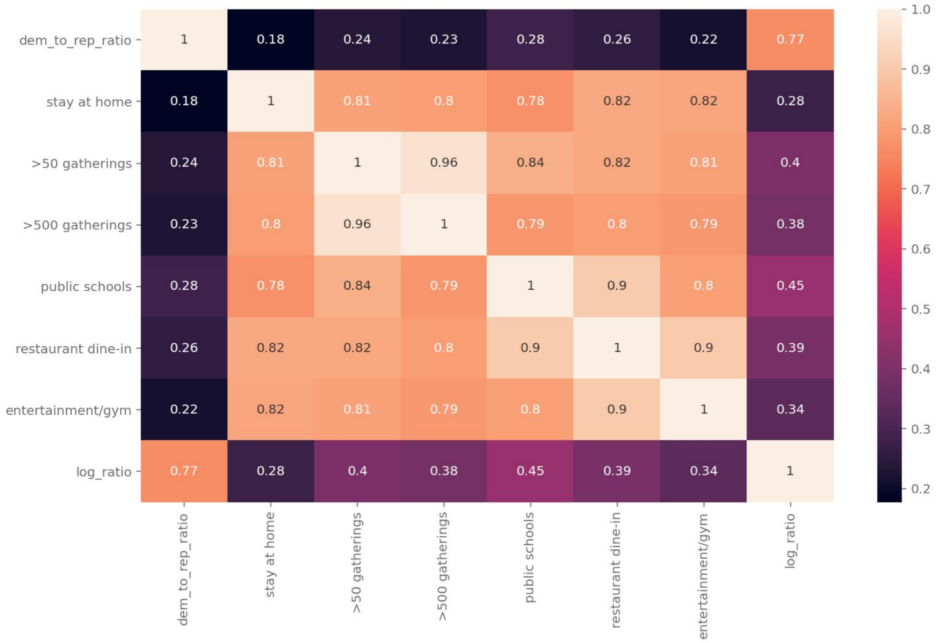

Fig 6. Correlation matrix for all 6 original covariates and `dem_to_rep_ratio` (untransformed) and its log transformation

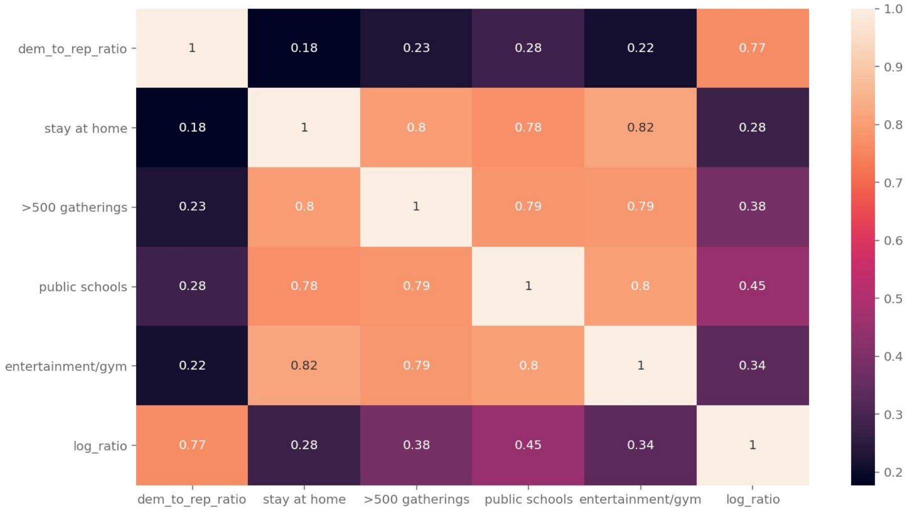

Fig 7. Correlation matrix for selected 4 covariates and `dem_to_rep_ratio` (untransformed) and its log transformation