

EDA: Characterizing the relationship between the religious affiliation and incidences of Covid-19 at U.S. universities.

Daniel de Castro and Laura Appleby

November 22, 2022

Description of data and source

Our data for this project comes from three sources:

1. NYTimes Covid-19 Data. This is publicly available on GitHub and was the source for some of the NYTimes maps and data visuals during the 2020-2021 era of the pandemic. It includes cases from 2020 - May 2021, and we have specifically selected cases at Universities. This dataset has 1948 entries and includes 2020 cases, 2021 cases, University IPEDS ID, University Name, State, etc.
2. IPEDS Data Center. This is publicly available data on colleges across the globe. It has many possible variables including demographics, admission rates, University affiliation, etc. The IPEDS data center allowed us to select certain Universities and variables. The smallest subset of Universities that included all from the NYTimes database (by IPEDS ID) was 6125 rows, with all US Universities.

The data is 1,855 rows after removing Universities without stats or without matching IPEDS ids. It has 40 columns, including IPEDS id, university name, cases, and predictor variables based on college attributes.

3. Centers for Disease Control. This publicly available data set tracks mask mandates in each state from April 8, 2020, to August 15, 2021.

Data Cleaning Procedures

For this exploratory data analysis, the first step is to read our data from CSV files into R data frames. The `colleges` data frame stores the NYT data on Covid cases at universities, while the `ipeds` data frame stores the data with most of our predictor variables (university characteristics) taken from IPEDS. We then rename most of the columns in `ipeds` to make them shorter and easier to work with.

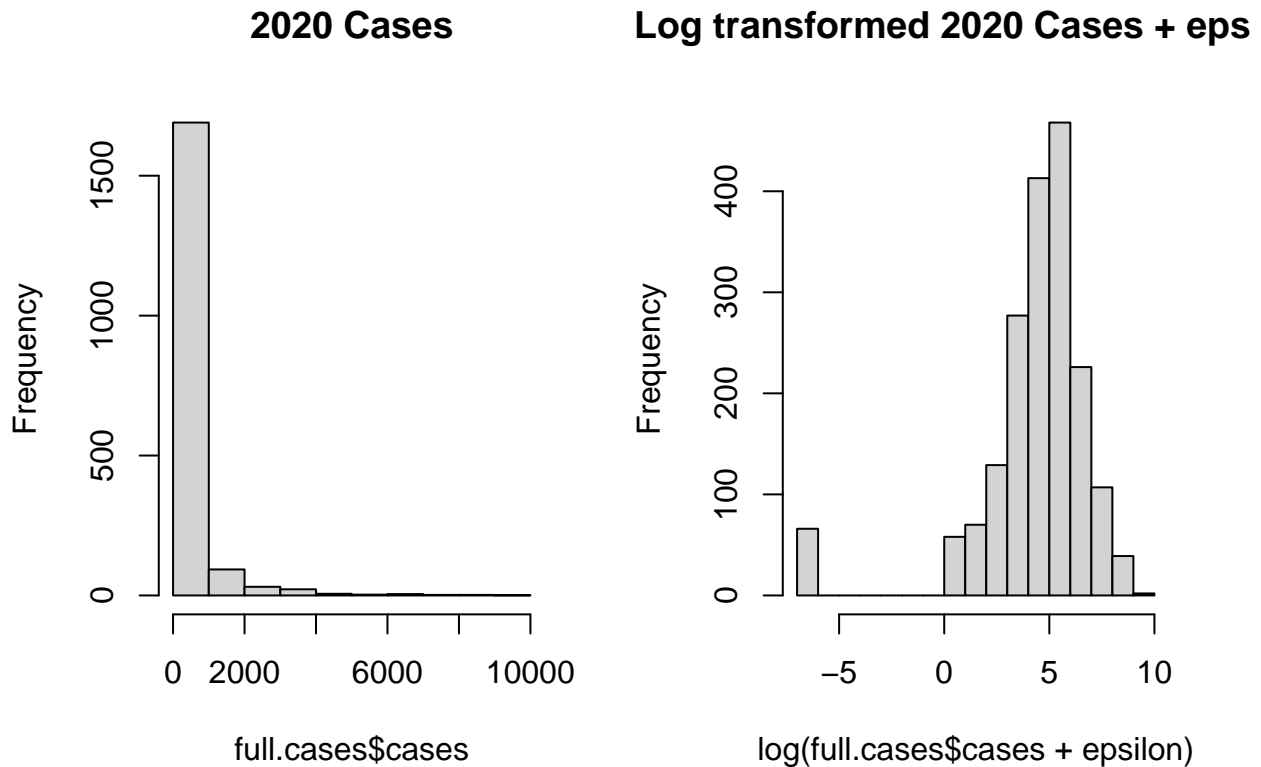
Next, we merge the `colleges` and `ipeds` data frames on the `ipeds_id` column and remove institutions with no IPEDS data. We then create the `religious`, `catholic`, and `private` columns, which are simply indicators for whether an institution has any religious affiliation, whether it has a catholic affiliation, and whether it is a private university. Finally, we drop the `control` column from the data frame, since it now contains redundant information.

We now look to add a column to the data frame that addresses the extent to which mask mandates were present in the state in which each institution is located. Below, we read out mask mandates data into a data frame from a CSV, treat the appropriate columns as factors, and convert `date` into R's `Date` type.

Now, we create a new simpler data frame to merge with `md`. This data frame contains only two columns: One with the name of each state, and the other with the number of days between July 1, 2020, and May 26, 2021, during which face masks were required in public in that state. We then merge this data frame with `md` to create `full.cases`, and finally create a column `total.cases` in `full.cases` that sums the `cases` and `cases_2021` columns.

EDA / Data Visualizations

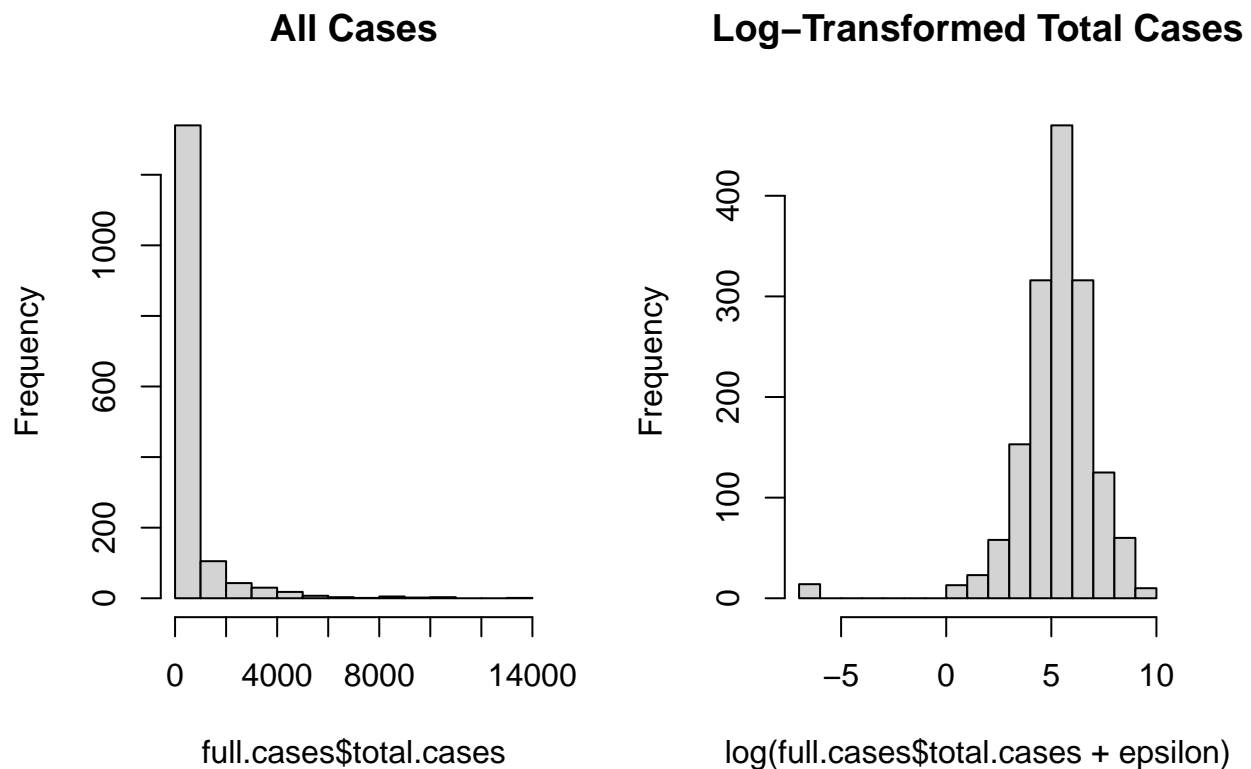
We will first look at the distribution of the total cases from the NYT data.



```
## [1] 39
```

From the above graphs we see that the un-transformed NYT case data from 2020 is strongly right skewed. In particular, there are 39 universities in the data set that recorded 0 Covid cases in 2020. Log transforming this data (and the inclusion of the arbitrary constant `epsilon = 0.001`) results in a more symmetrically distributed distribution with some negative outliers on the left. These outliers correspond to those observations with zero Covid cases in 2020.

We will now have a look at the distribution of `total.cases` below.



```
## [1] 39
```

There are still 39 institutions with no reported Covid cases. The resulting plots are nearly identical to the plots for only the cases reported in 2020. The main difference between the two sets of plots is that the magnitude of the observations in `total.cases` is generally larger, which makes sense given that `total.cases` also includes case counts from the first five months of 2021. Thus, the histogram of the log-transformed distribution of `total.cases` is shifted slightly to the right compared to the log-transformed distribution of `cases`.

Given that we are focusing on the relationship between the religious affiliation of an institution and the number of Covid cases it reported in the 2020-21 academic year, we will also find it useful to look at a boxplot comparing case counts between religious and non-religious institutions:

Total cases in total headcount by binary religious affiliation



```
##
##   No   Yes
## 1372 483
```

Above we see that the median number of cases as a fraction of total headcount (to account for differing school sizes) at religious universities is slightly higher than that of non-religious schools, and the 75th quartile much higher. Thus, there does seem to be some sort of difference in terms of case counts between the groups. In our actual project, we could formalize this observation by performing a parametric test for a difference in sample means, or some sort of non-parametric test (such as a rank-based test). Both distributions also appear to be somewhat right-skewed, but the distribution for non-religious institutions possibly is more so. This is likely because there are more non-religious institutions in the dataset (1372 vs 483 religious) and they likely represent a more diverse pool: non-religious schools can be large public institutions (ie. The UC schools) or small private institutions (ie. Wesleyan). Religious schools are always private based on IPEDS classifications.

Baseline Model

To establish that our predictors would have a relationship with `total.cases`, we fit a simple linear model that predicts `total.cases` from the predictors that we are the most interested in.

```
##
## Call:
## lm(formula = log(total.cases + epsilon) ~ religious + FIPS.state.code +
##     total.headcount + undergrad.headcount + percent.american.native +
##     percent.asian + percent.black + percent.hispanic.latino +
##     percent.pacific.islander + percent.white + percent.two.more.races +
##     percent.NA.race + percent.disability + catholic + dorm.capacity +
##     private + mask.mandated.days, data = full.cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4839  -0.3591   0.1022   0.5268   2.0884
```

```

##
## Coefficients: (1 not defined because of singularities)
##
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.329e+00 7.470e-01 7.134 2.25e-12 ***
## religiousYes 7.737e-02 1.088e-01 0.711 0.477397
## FIPS.state.codeAlaska -2.364e+00 1.040e+00 -2.273 0.023324 *
## FIPS.state.codeArizona -1.002e+00 5.631e-01 -1.780 0.075463 .
## FIPS.state.codeArkansas 1.090e-01 4.041e-01 0.270 0.787426
## FIPS.state.codeCalifornia -1.267e+00 3.651e-01 -3.472 0.000546 ***
## FIPS.state.codeColorado 1.750e-01 4.027e-01 0.435 0.663915
## FIPS.state.codeConnecticut 6.864e-02 3.738e-01 0.184 0.854345
## FIPS.state.codeDelaware 2.450e-01 6.048e-01 0.405 0.685556
## FIPS.state.codeDistrict of Columbia -5.084e-01 5.032e-01 -1.010 0.312673
## FIPS.state.codeFlorida -8.978e-02 3.522e-01 -0.255 0.798830
## FIPS.state.codeGeorgia -2.601e-02 3.587e-01 -0.072 0.942226
## FIPS.state.codeHawaii -9.086e-01 1.120e+00 -0.811 0.417664
## FIPS.state.codeIdaho 1.302e-01 5.569e-01 0.234 0.815156
## FIPS.state.codeIllinois -2.984e-01 3.309e-01 -0.902 0.367519
## FIPS.state.codeIndiana -4.909e-01 3.545e-01 -1.385 0.166508
## FIPS.state.codeIowa -5.817e-01 3.810e-01 -1.527 0.127164
## FIPS.state.codeKansas -3.469e-01 4.995e-01 -0.695 0.487564
## FIPS.state.codeKentucky -3.945e-01 3.540e-01 -1.114 0.265431
## FIPS.state.codeLouisiana -4.320e-01 4.023e-01 -1.074 0.283342
## FIPS.state.codeMaine -6.353e-01 4.350e-01 -1.461 0.144539
## FIPS.state.codeMaryland -4.506e-01 3.806e-01 -1.184 0.236731
## FIPS.state.codeMassachusetts -8.964e-01 3.206e-01 -2.796 0.005304 **
## FIPS.state.codeMichigan -3.232e-01 3.361e-01 -0.962 0.336582
## FIPS.state.codeMinnesota -3.176e-01 3.541e-01 -0.897 0.370037
## FIPS.state.codeMississippi -1.270e-01 4.993e-01 -0.254 0.799330
## FIPS.state.codeMissouri -2.952e-01 3.582e-01 -0.824 0.410070
## FIPS.state.codeMontana -1.028e-01 6.057e-01 -0.170 0.865208
## FIPS.state.codeNebraska -3.362e-02 4.476e-01 -0.075 0.940145
## FIPS.state.codeNevada 6.058e-01 7.306e-01 0.829 0.407228
## FIPS.state.codeNew Hampshire -2.209e-01 4.341e-01 -0.509 0.610990
## FIPS.state.codeNew Jersey -2.156e-01 3.517e-01 -0.613 0.539939
## FIPS.state.codeNew Mexico 1.727e-01 9.828e-01 0.176 0.860539
## FIPS.state.codeNew York -5.616e-01 3.008e-01 -1.867 0.062260 .
## FIPS.state.codeNorth Carolina -2.245e-01 3.342e-01 -0.672 0.501821
## FIPS.state.codeNorth Dakota -5.545e-01 6.032e-01 -0.919 0.358282
## FIPS.state.codeOhio -4.964e-01 3.231e-01 -1.536 0.124834
## FIPS.state.codeOklahoma -1.006e-02 6.220e-01 -0.016 0.987095
## FIPS.state.codeOregon -1.092e+00 3.981e-01 -2.744 0.006220 **
## FIPS.state.codePennsylvania -4.945e-01 3.026e-01 -1.634 0.102679
## FIPS.state.codeRhode Island 1.374e-01 4.354e-01 0.316 0.752428
## FIPS.state.codeSouth Carolina 3.377e-02 3.661e-01 0.092 0.926530
## FIPS.state.codeSouth Dakota -1.319e-01 9.619e-01 -0.137 0.890949
## FIPS.state.codeTennessee -1.500e-01 3.791e-01 -0.396 0.692392
## FIPS.state.codeTexas -3.538e-02 3.525e-01 -0.100 0.920077
## FIPS.state.codeUtah 3.805e-01 5.434e-01 0.700 0.484023
## FIPS.state.codeVermont -1.258e+00 4.610e-01 -2.730 0.006482 **
## FIPS.state.codeVirginia -3.001e-01 3.376e-01 -0.889 0.374352
## FIPS.state.codeWashington -1.238e+00 3.766e-01 -3.286 0.001060 **
## FIPS.state.codeWest Virginia -4.189e-01 4.318e-01 -0.970 0.332239
## FIPS.state.codeWisconsin -5.312e-01 3.445e-01 -1.542 0.123503

```

```
## FIPS.state.codeWyoming      -3.383e-01  7.095e-01  -0.477  0.633591
## total.headcount             2.866e-05  1.465e-05   1.955  0.050894 .
## undergrad.headcount         2.177e-06  1.771e-05   0.123  0.902186
## percent.american.native     -6.335e-03  2.777e-02  -0.228  0.819597
## percent.asian               -8.831e-03  1.393e-02  -0.634  0.526169
## percent.black               -1.179e-02  7.767e-03  -1.518  0.129424
## percent.hispanic.latino     -6.108e-03  8.642e-03  -0.707  0.479919
## percent.pacific.islander     1.340e-02  1.275e-01   0.105  0.916358
## percent.white               9.443e-03  6.961e-03   1.357  0.175319
## percent.two.more.races      -4.699e-02  2.290e-02  -2.051  0.040565 *
## percent.NA.race             -1.746e-02  8.493e-03  -2.056  0.040154 *
## percent.disability          -1.393e-02  6.090e-03  -2.288  0.022416 *
## catholicYes                 2.441e-01  1.209e-01   2.019  0.043793 *
## dorm.capacity               2.555e-04  2.348e-05  10.885  < 2e-16 ***
## privateYes                  -2.779e-02  1.096e-01  -0.254  0.799876
## mask.mandated.days          NA          NA          NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9153 on 769 degrees of freedom
## (1020 observations deleted due to missingness)
## Multiple R-squared:  0.6274, Adjusted R-squared:  0.5959
## F-statistic: 19.92 on 65 and 769 DF,  p-value: < 2.2e-16
```

As we can see from the R output above, a handful of these predictors were found to be significant at the $\alpha = 0.05$ level in this model, including `catholic`, `percent.disability`, and `dorm.capacity`. Importantly, `catholic` addresses our hypothesis as much as the predictor `religious`, since it makes a more specific claim about the religious affiliation about a particular institution. In this model, we find that `catholic` has a positive association with log-transformed `total.cases`. We also find that the coefficient estimates for `percent.disability` and `dorm.capacity` are negative, which is somewhat surprising: Our gut feeling would be to expect both of these predictors to have positive associations with `total.cases`, since having more immunocompromised students on campus (`disabled` is a somewhat flawed measure of this predictor) and having more students living in dorms on campus might be thought to increase case counts. It could be the case, however, that having more immunocompromised students and more students living on campus might prompt a university to take precautions against Covid more seriously.

It is also interesting that the model failed to produce coefficient estimates for `mask.mandated.days`. This is likely due to strong — indeed, almost perfect — collinearity between this predictor and `FIPS.state.code`, since each value of `FIPS.state.code` does indeed correspond to some specific and unchanging value in `mask.mandated.days`. Thus, it appears that, in the models that we will prepare for the final paper, we either need to include a more nuanced predictor variable reflecting mask mandates at each institution, or decide that we are happy with the representation of the presence of mask mandates at each school that will be captured by a predictor such as state or county. To that end, performing an ANOVA to test whether the mean of `total.cases` is different for institutions in each state or even each county might be an interesting exercise as we put together this final paper.