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

Data comes from two sources:

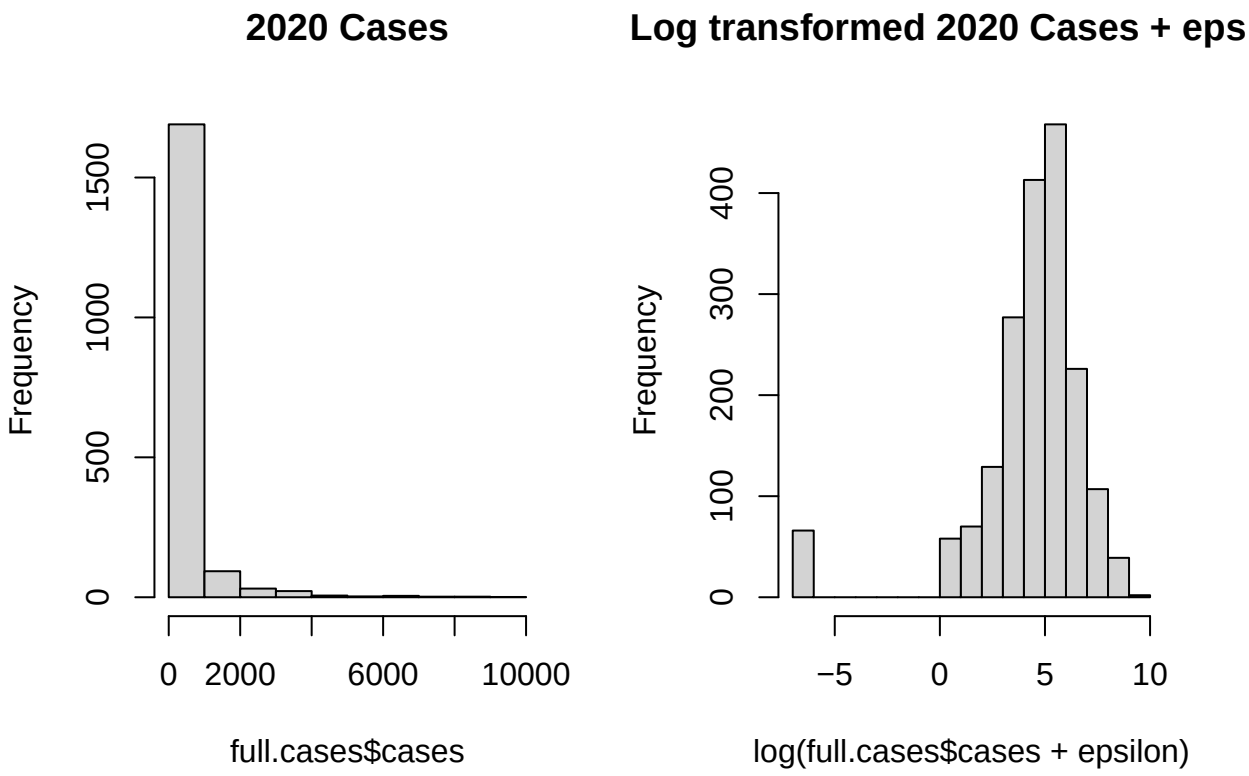
1. NYTimes Covid-19 Data. This is publicly available on GitHub and was the source for some of the NYTimes maps and data visuals during the 2020-2021 era of the pandemic. It includes cases from 2020 - May 2021, and we have specifically selected cases at Universities. This dataset has 1948 entries and 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.

EDA / Visuals of data

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 will 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.

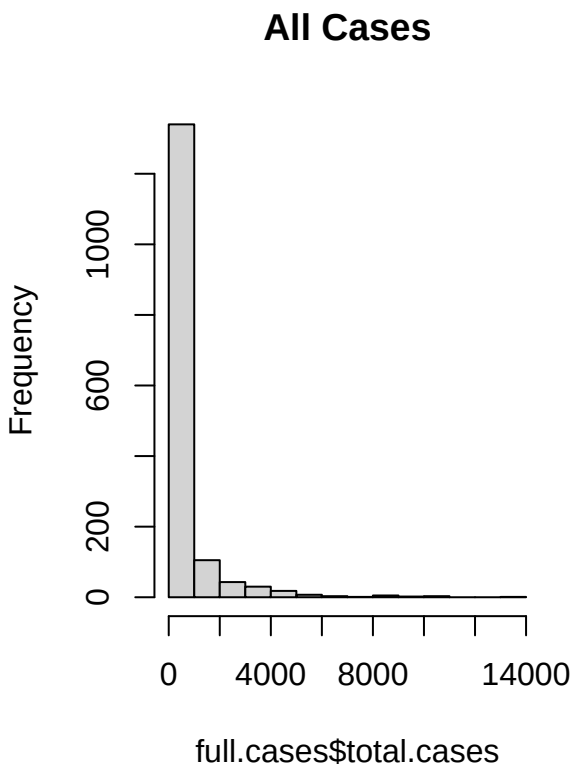
We will first look at the disribution of the cases from the NYTimes data.

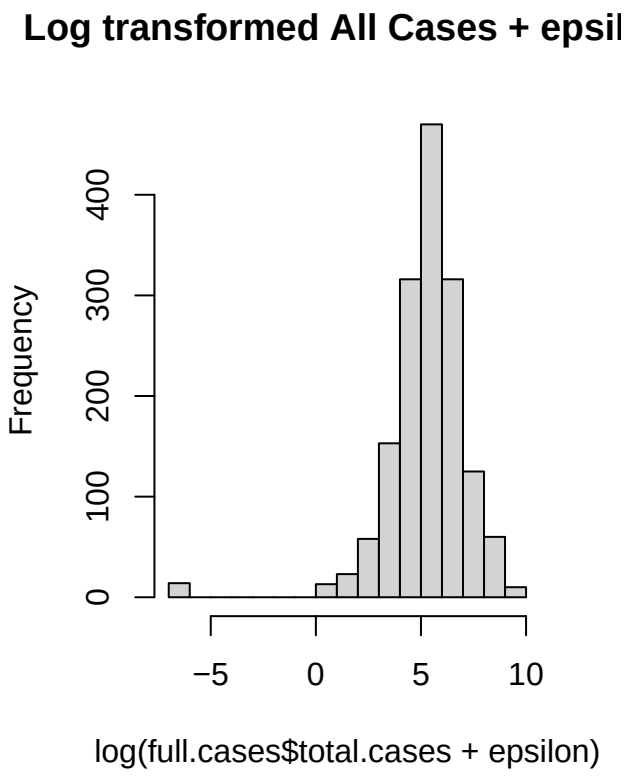



```
## [1] 38
```

From the above graphs we see that the un-transformed NYTimes case data from 2020 is strongly right skewed. In particular, there are 38 universities in the dataset that recorded 0 Covid cases in 2020. Log transforming this data (+ arbitrary epsilon of 0.001) results in a more normally distributed data with some negative outliers on the left. This is from the 38 Universities that reported no cases.

We will now look at total cases (for 2020 and 2021), below.

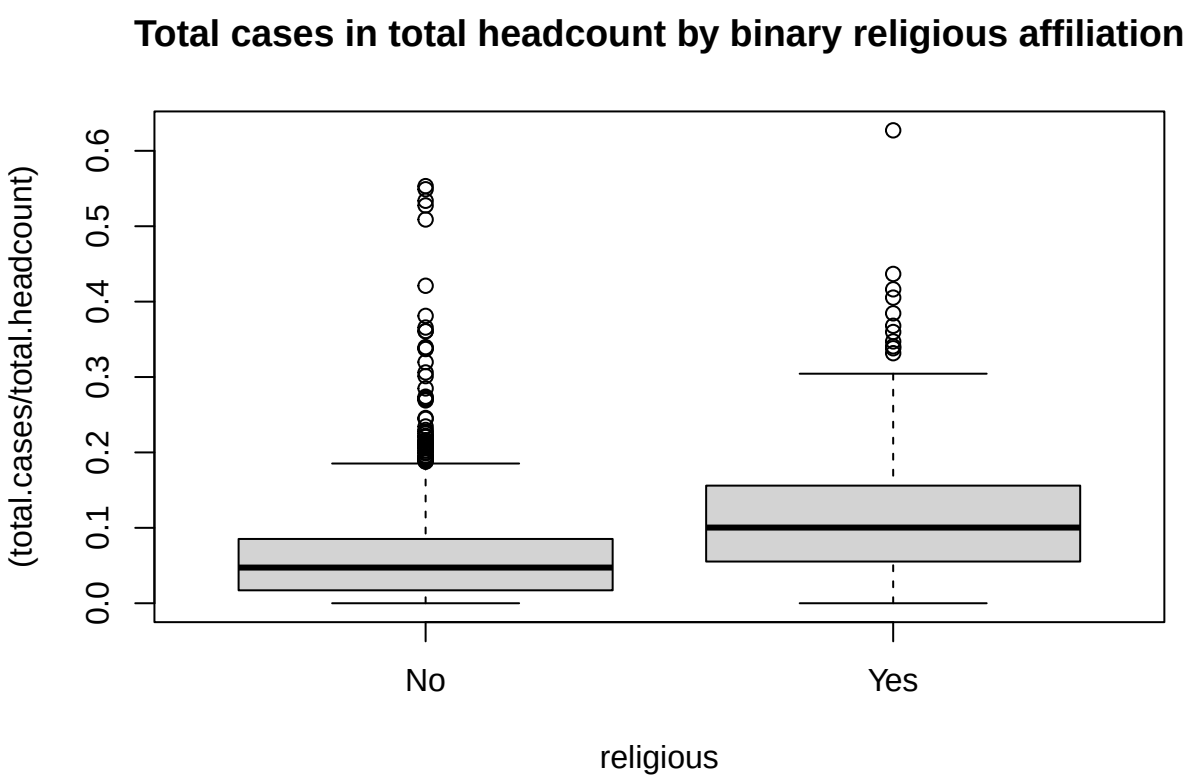





```
## [1] 38
```

There are still 38 institutions with no reported covid cases. The resulting plots are nearly identical to the 2020-only plots above. The main difference is that there are simply some more cases reported, extending the lines of the histogram vertically when on the same scale.

Given that we are focusing on the relationship between the religious affliction of an institution and its 2020/2021 Covid cases, we will also look at a simple boxplot quantile breakdown between religious and non-religious institutions:




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

Above we see that (when accounting for school size via `total.headcount`) the median number of cases (by percent) at religious universities slightly higher than that of non-religious schools and the 75th quartile much higher. But, the non-religious schools have a more visible right spread. 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-regligious 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 for reference

We explored several baseline models. Below we compared religious, control and tuition to the total cases from 2020 and 2021.

```
##
## Call:
## lm(formula = log(total.cases + epsilon) ~ religious + control +
##     tuition, data = full.cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2124  -0.7662   0.1396   0.9542   5.3797
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.457e+00  3.957e-01   3.681 0.000241 ***
## religiousYes                  7.047e-01  1.400e-01   5.035 5.37e-07 ***
## controlPrivate not-for-profit 8.779e-01  4.185e-01   2.098 0.036108 *
## controlPublic                 3.519e+00  3.916e-01   8.986  < 2e-16 ***
## tuition                       6.641e-05  5.433e-06  12.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.668 on 1470 degrees of freedom
##   (380 observations deleted due to missingness)
## Multiple R-squared:  0.1219, Adjusted R-squared:  0.1195
## F-statistic: 51.01 on 4 and 1470 DF,  p-value: < 2.2e-16
```

It appears that all variables have a positive relationship with cases, to varying degrees. Relationships exist between variables such as private not-for-profit universities all being non-religious mean that it is excluded on account of multicollinearity.

We will explore these relationships further, as well as add-onto our baseline model and explore other models.