Final Report W203 - Team 2

Stephen Chen

7/31/2021

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

## `summarise()` has grouped output by 'Recip_State'. You can override using the `.groups` argument.

## Joining, by = "State"

## NULL

## `summarise()` has grouped output by 'state'. You can override using the `.groups` argument.

## Joining, by = "State_full"

## NULL

## NULL

## NULL

## NULL

## Joining, by = "State"
## ‘summarise()‘ has grouped output by ’State.Postal.Code’. You can override using the ‘.groups‘ argumen
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
##
## Call:
## lm(formula = total_vacation_trips_prop ~ Vaccination_rate_prop,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70137 -0.21465 -0.04035  0.07984  1.21023
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.1564     0.2639    8.17 1.66e-10 ***
## Vaccination_rate_prop -278.1314    70.9469   -3.92 0.000292 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3918 on 46 degrees of freedom
## Multiple R-squared:  0.2504, Adjusted R-squared:  0.2341
## F-statistic: 15.37 on 1 and 46 DF,  p-value: 0.0002922
```

```
##            Vaccination_rate_prop           log10(Cases_rate_prop)
##                         1.948908                         1.063781
##     log10(stay_at_home_prop + 1)          log10(Deaths_rate_prop)
##                         1.445748                         1.313793
## total_mask_mandate_duration_prop                 Weekly_UI_Amount
##                         1.973961                         1.545134
```

```
##
## Call:
## lm(formula = total_vacation_trips_prop ~ Vaccination_rate_prop +
##     log10(Cases_rate_prop) + log10(stay_at_home_prop + 1) + log10(Deaths_rate_prop) +
##     total_mask_mandate_duration_prop + Weekly_UI_Amount, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63847 -0.26878 -0.07505  0.14009  1.12311
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.205e+00  2.561e+00  -0.861  0.39428
## Vaccination_rate_prop           -3.270e+02  9.822e+01  -3.330  0.00185 **
## log10(Cases_rate_prop)          -1.506e-01  5.086e-01  -0.296  0.76862
## log10(stay_at_home_prop + 1)    -1.052e+00  2.273e+00  -0.463  0.64595
## log10(Deaths_rate_prop)         -6.361e-01  3.582e-01  -1.776  0.08315 .
## total_mask_mandate_duration_prop -1.313e-02  2.704e-01  -0.049  0.96151
## Weekly_UI_Amount                 2.903e-04  4.792e-04   0.606  0.54798
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 0.3885 on 41 degrees of freedom
## Multiple R-squared:  0.343,  Adjusted R-squared:  0.2469
## F-statistic: 3.568 on 6 and 41 DF,  p-value: 0.006176
```

```
##
## ====================================================================================================
##                                                         total_vacation_trips_prop
##                                             (1)                    (2)                   (3)
## ----------------------------------------------------------------------------------------------------
## log10(Vaccination_total)                                                                -25.460
##                                                                                         (16.070)
##
## log10(Cases_total)                                                                     11.966***
##                                                                                         (3.558)
##
## log10(Deaths_total)                                                                     -2.933*
##                                                                                         (1.507)
##
## log10(Vaccination_rate)                                                                 22.900*
##                                                                                         (13.322)
##
## log10(Cases_rate)                                                                        -4.478
##                                                                                         (2.856)
##
## log10(Deaths_rate)                                                                       -1.963
##                                                                                         (1.716)
##
## Vaccination_total_prop                                                                    29.302
##                                                                                         (17.579)
##
## Cases_total_prop                                                                       -52.027***
##                                                                                         (18.542)
##
## Deaths_total_prop                                                                      635.746
##                                                                                        (413.031)
##
## Vaccination_rate_prop                     -278.131***            -327.033***           -3,210.048*
##                                             (70.947)               (98.216)            (1,608.321)
##
## log10(Cases_rate_prop)                                             -0.151
##                                                                    (0.509)
##
## Cases_rate_prop                                                                        13,912.620
##                                                                                       (9,127.642)
##
## Deaths_rate_prop                                                                      245,438.900
##                                                                                      (312,645.800)
##
## log10(stay_at_home_prop + 1)                                       -1.052                -2.646
##                                                                    (2.273)               (2.367)
##
## log10(Deaths_rate_prop)                                            -0.636*
##                                                                    (0.358)
##
```

```
## eviction_moratorium_duration_prop                                                      0.292
##                                                                                       (0.220)
##
## total_mask_mandate_duration_prop                                -0.013                 -0.129
##                                                                (0.270)                (0.268)
##
## Weekly_UI_Amount                                                0.0003                 0.0002
##                                                                (0.0005)               (0.0004)
##
## Constant                                 2.156***               -2.205                 23.485
##                                          (0.264)               (2.561)                (30.971)
##
## N                                           48                    48                     48
## R2                                         0.250                 0.343                  0.712
## Adjusted R2                                0.234                 0.247                  0.563
## Residual Std. Error                   0.392 (df = 46)        0.389 (df = 41)       0.296 (df = 31)
## F Statistic                        15.369*** (df = 1; 46) 3.568*** (df = 6; 41) 4.789*** (df = 16; 31
## ====================================================================================================
## Notes:                                                          ***Significant at the 1 percent level
##                                                                  **Significant at the 5 percent level
##                                                                   *Significant at the 10 percent level
```

Introduction

Tourism is one of the biggest industries in the United States. In 2019, it supported 15.8 million jobs and generated $2.6 billion dollars. However, the COVID-19 pandemic disrupted what would have been lucrative spring and summer 2020 vacation seasons, with most of the country encouraged to socially distance, wear mask, and stay at home as much as possible. Because COVID-19 spreads very easily via respiratory droplets and infect victims asymptomatically, many people sacrificed the desire to congregate in large groups in order to keep their loved ones safe, especially those who were high-risk. Thus, leisure and travel plummeted in 2020. Travel spending in 2020 fell a record 42% from 2019 and resulted in $426 billion in cumulative losses to the travel economy.

https://www.ustravel.org/toolkit/covid-19-travel-industry-research

Government mandates and the public's encouraged adherence to them were to help lower the number of cases and deaths caused by COVID-19 so that the disease would not spread rapidly while the vaccine was being created and approved. And with the creation of the COVID-19 vaccines, we've seen a high demand for the vaccine in the beginning and still continue to see a majority of the population in the United States who choose to gain some semblance of an immunity against COVID-19. As such, government mandates meant to counteract the spread of COVID-19 began to lift. Because of these decisions and the population's general anxiousness to return to a "normal" life, popular destinations for leisure and vacation like restaurants and theme parks began to open and support a larger capacity.

However, state governments and the constituents all chose different ways in handling the pandemic and subsequently felt different negative economic impacts that accrued. During the pandemic, some states issued mask mandates and stay at home orders, while other states didn't. Due to businesses seeing less income, many companies laid off their workers in the interest of saving costs or the job just couldn't be adjusted to function remotely, leading to some of the population's incomes becoming destabilized. To counteract this, some states issued eviction moratoriums and unemployment income to stimulate the economy and protect tenants going through economic turmoil. In addition, not all of the population have taken the vaccine, due to personal beliefs or limited supply.

The goal of our project is to research the effect that the COVID-19 pandemic had on the number of vacations taken. More specifically, we want to look at how states' performance and actions against the pandemic affected their constituent's propensity to travel. We would want to find if a state's characteristics and performance against COVID-19 changed the way its population traveled for leisure and/or vacation, or even if those variables even mattered in the way the population wanted to travel. Our dependent variable of interest here is the number of trips on average for leisure and/or vacation per capita. Here, we are assuming that a trip traveled greater than 250 miles in any mode of transportation. We are also limiting the timeframe to the Spring 2021 season (March 1st, 2021 to May 31st, 2021) to account for the widespread availability of the COVID-19 vaccine and state governments' decision to lift mandates meant to limit the spread of COVID-19.

The independent variables that we are focusing on are areas related to the states' performances against COVID-19, like cases and deaths caused by the virus. We are also interested in the states' policies to counteract COVID-19, such as stay at home mandates, mask mandates, and unemployment income. Lastly, we also looked at state population vaccination against COVID-19.

Data Description

Our data sources include the COVID-19 US state policy database (CUSP), the New York Times cumulative counts of COVID-19 cases and deaths at the state level, the Bureau of Transportation Statistics (BTS) Trips by Distance dataset, and the Centers for Disease Control and Prevention (CDC) County Level COVID-19 Vaccination Dataset.

Vaccination Data

The CDC County Level COVID-19 Vaccination Dataset was also lengthy due to county level granularity, at approximately 778K rows; of these we only made use of one, (the Recip_State for grouping and Date which was aggregated were also used) Series_Complete_Yes. Series_Complete_Yes being the total number of people who are fully vaccinated (have a second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where the recipient lives.

Data Manipulation

The first step was to only include vaccination data that fit our Spring 2021 timeframe, filtering the data from March 1st, 2021 to May 31st, 2021. The only variable we were interested was the data that showed the number of fully vaccinated persons. Since the dataset was recorded by county, we aggregated together the total vaccinated members grouped by state and day. [They] recorded these numbers as a running total. Thus, to get a rate of the number of people fully vaccinated people per day, we subtracted the previous day by the next to get the number of new persons vaccinated per day. Finally, we averaged all of the daily rates in the Spring 2021 timeframe to determine the state's average rate of fully vaccinated persons per day. We treated as the total number of fully vaccinated persons at May 31st, 2021 as the total number of fully vaccinated persons. To note, Hawaii and Texas were the two states in this dataset that didn't have any vaccination data. We had to remove these two states from our final dataset in the interest of having a clean dataset. To balance the total population of fully vaccinated persons against the differing populations in each state, we divided the total number of fully vaccinated persons by the the state's population in 2018 to get a proportion of the total population in each state that is fully vaccinated. To balance the rate of new fully vaccinated persons per day, we also divided the rate by the state's population in 2018.

State Performance against COVID-19

The New York Times cumulative counts of COVID-19 cases and deaths at the state level was the simplest of datasets, containing only the date, state, fips number, cases, and deaths for each state over time. We used the date, state, cases, and deaths which were all further manipulated before the variables were introduced into our models.

Data Manipulation

The first step was to only include data that fit our Spring 2021 timeframe, filtering the data from March 1st, 2021 to May 31st, 2021. The two variables of interest in this dataset were the number of cases and deaths caused by COVID-19. This dataset recorded these numbers as running totals, so to get the rate of number of cases and deaths caused by COVID-19 per day, we subtracted the previous day by the next to get the number of new cases and deaths caused by COVID-19 per day. We then averaged all of the daily rates in the Spring 2021 timeframe to determine the state's average rate of new cases and deaths caused by COVID-19 per day. To balance the totals and the rates against the differing populations in each state, we divided these numbers by the state's population.

State Characteristics and Government Policies/Mandates against COVID-19

The CUSP database contained multiple datasets which were used in the analysis. From the CUSP database we used State Characteristics, Face Masks, Stay at Home, Housing, and Unemployment Insurance. From State Characteristics we pulled only the population. From the Face Masks dataset we used the start date and end date fields for the face mask mandate. From the Stay at Home dataset we used the start date and end dates of the stay at home mandate. From the State Housing dataset we used the start and end dates of the eviction moratoriums.

Data Manipulation

A great deal of manipulation was involved to make the data from the CUSP database more straightforward and more compatible with our analysis. In each case the durations of multiple mandate policy periods were summed over what we determined as the Covid Era. The Covid Era is the period between Jan 20th 2020 (the date of the first documented case of COVID-19 in the US) and August 2nd (the date we determined this metric). Dividing by the duration of the Covid Era transforms the mandate duration values (days) into more easily comparable values (proportion of covid_era) as they are all restricted between 0 and 1. Lastly in the case of the Stay at Home dataset, the field for the stay at home orders which were issued but did not specifically restrict movement of the general public were re-categorized as simply stay at home orders. We chose to include these values as without inclusion more states would not have any stay at home mandate durations which we felt was not truly representative of their state policy.

State Trips by Distance

The BTS Trips by Distance dataset is lengthy at approximately 3M rows, however this is due to the county level granularity of the dataset, filtering off for state level data we only made use of two fields, number of trips between 250 and 500 miles, and the number of trips greater than 500 miles. The reasoning behind this is motivated by the National Household Travel Survey done by the US Department of Transportation (https://nhts.ornl.gov/briefs/Vacation%20Travel.pdf) which states "Long distance vacation trips by car… [are] an average of 314 miles one-way."

Data Manipulation

The first step was to only include data that fit our Spring 2021 timeframe, filtering the data from March 1st, 2021 to May 31st, 2021. The two variables of interest in this dataset were the number of trips taken that were 250 - 500 miles and the number of trips taken that were 500 and above. Sorting by the day and state, we added these two numbers together, based on our decision that trips 250 miles and greater were vacation trips. Finally, we divided the total number of vacation trips in the Spring 2021 timeframe by the state's population in 2018 to determine the state's vacation trips taken per capita.

The Model Building Process

ERICK: The first model you include should include only the key variables you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (perhaps one, or at most two-three, covariates if they are so crucial that it would be unreasonable to omit them).
Our limited model:
(insert latex here) This model is very simple, but is a direct result of analysis of various models which found little to no statistical significance when vaccination rate or collinear variables to vaccination rate were included in the model. To begin, we included every variable that would be reasonable to include and which would be comparable between states such that there would be no bias toward one state or another based on the various populations. Such potentially problematic variables include (death count, covid count, vaccination count, etc). The regressions on such inclusive models indicated only a few significant variables. Limiting the variables to include only those with the greatest magnitude in coefficients, and which were not found to be collinear, returned vaccination rate, COVID-19 case rate, and death rate. However, the only variable with consistent statistically significant coefficients in various iterations (even in models with various state policies) would be the vaccination rate. It seems public policy does not have great correlation with the number of vacation trips, i.e. people tend to take vacation trips based much more on their state's vaccination rates than their state's public policy. Investigating the relationship between Vaccination_rate_prop and total_vacation_trips_prop more closely we find the following:

(insert stargazer)

The negative correlation between Vaccination_rate_prop (average vaccination rate per capita) and the total_vacation_trips_prop (total number of trips greater than 250 miles per capita) can be rationalized with consideration to the heavily politicized landscape of the US as it stands. States which have a high average vaccination rate per capita would also likely have a greater proportion of left-leaning citizens, which would likely trend negatively with a great number of vacation trips during COVID-19. Likewise, states with a low average vaccination rate per capita would also likely have a greater proportion of right-leaning citizens, which would likely trend positively with a great number of vacation trips during COVID-19 due to conservative values of freedom and liberty. (See omitted variable bias section for further discussion on political leaning.) Our second model looks into the trends that affect that of the proportion of U.S. citizens and their vacationing habits as COVID-19 has taken a grip on the nation. The goal of this model is to explain the effects that each has on the proportion of vacations taken by the U.S. The covariates used for our second model are vaccination total proportion, vaccination rate proportion, case rate proportion, death rate proportion, vacation trips taken proportion, and stay at home order proportion and the weekly amount of unemployment checks per state.

We model the relationship between the individual state policies, if such states adhered to that policy at all, as each state was not under federal mandate to employ such a policy, this allowing for independence amongst states and how they would decide what is best for the overall health of the area. Each of the proportions gives us a more detailed indication of whether vacation decisions were positively trending or not. Also, we take a look into seeing if the proportion of cases per state or the proportion of deaths per state would influence travellers to either stay at home and travel less or feel more emboldened to travel more. As we see in our model here (model image to be above) we see that there are generally negative trends in the amount of vacations taken. We observe that our statistically significant predictors in this more detailed model are our vaccination rate proportion, the log of the death rate proportion. Our model here is not highly correlated with all covariates involved remaining under a rating of two within a VIF test. This model shows even further that vacations are taken at a lesser rate when including the factors of, mask mandate policy, stay at home orders, death rates, and case rates, giving the indication that with more people being vaccinated, it is possible that more people are aware of the virus and are less likely to go on vacations as often as they would if they weren't vaccinated.

Our third model included covariates in our second model and all other covariates in our dataset. It looks into all of the possible trends captured by our dataset where the pandemic affected the vacation habits of people in the United States. Beyond the weekly unemployment income, stay at home mandate length proportion, and the mask mandate length proportion, it also includes the eviction moratorium duration proportion. It also includes all of the rate per day and total fully vaccinated persons, COVID-19 cases, and COVID-19 deaths, as well as their proportion to the state's population. Because the total number and rates of vaccinated persons and cases/deaths caused by COVID-19 and the stay at home duration proportion were heavily skewed, these covariates needed to be log-transformed. This resulted, unsurprisingly, in the highest R^2 of all of the models. However, the only statistically significant predictors were the log of the total COVID-19 cases and the total COVID-19 cases in proportion to the state population. It was also easy to see that this model would suffer from multicollinearity issues. In fact, the only covariates that did not show any collinearity were the state government mandates and weekly unemployment income. This makes sense because many of our vaccination and COVID-19 performance variables were manipulated and derived using similar processes. Thus, those covariates ended up being highly correlated with each other, resulting in collinearity issues. This not only impacts the statistical significance of the coefficients from the model, but it also impacts the practical significance of the coefficients produced by the model. This model is not recommended for practical use because of this glaring issue. To resolve it, we will have to carefully choose which covariates to include in the model; we would ideally only want to choose one covariate each that describes vaccinations, performance metrics against COVID-19, and state government's use of policies counteracting COVID-19.

Limitations of Our Models

The limitations of this model include heavy manipulation of the data sets in order to gain balance amongst the states to evaluate statistical significance and normal distribution with respect to specific variables such as vaccination totals, which are much larger than other states due to their population sizes or individual state requirements. We needed to manipulate and transform every variable that was only a total to become a proportion to better grasp the impact of specific totaled variables and measure them in our regression properly.

We also encountered issues with states not recording data such as Texas and Hawaii not recording their vaccination data at all. This caused disruption within our models and would skew data incorrectly. However, with Texas being such a populated state, one of the most populated in the contiguous U.S, the removal of such a state would also negatively impact our regression, affecting the significance of all of our covariates as Texas' death rate (61.5) was the second highest ranking state behind California, with a 78.8 death rate, and a case rate of 3087.96 which was our third highest amongst all states, only behind Florida and New York.

Another limitation that was identified was that of geographic location and misleading values for specific states, like that of Alaska, which geographically is not connected to the lower 48 states and therefore the majority of vacations for this state would be over 250, generally qualifying most travel as a vacation just to get to the closest state. We found it necessary to remove D.C from our dataset as well considering the fact that it was not a state, but a district, to keep all elements relevant to state to state travel in this exploration of vacation trips on a state by state basis.

Outliers in our data also proved to be problematic as we identified states like California having the most amount of vaccinations in the entire data set by at least 8 million and outliers within a small time frame proved to affect the modeling process greatly, causing the need to transform vaccination, case, death, and vacation total and rates to proportions as not only is the population much greater than that of most other states, but the overall amounts from this state significantly greater than many other states observed, including the longest stay at home mandate and mask mandate orders in the country.

Discussion of Omitted Variables

Our Limited Model: (insert latex here) Considering this relationship some reasonable omitted variables not previously addressed persist. The most obvious might be income. Vacationing is an activity which is highly affected by the amount of disposable income available to the individual especially given our narrow definition for vacation. It takes a significant amount of money to travel 250+ miles, by any means (car, plane, boat, train, etc). We tried to incorporate some aspect of this variable using the Unemployment Insurance data, however the regression coefficient was small (has small impact in the model relative to the other variables) and it was not statistically significant. Incorporating an input variable specifically detailing the amount of disposable income associated per person of each state would likely be an important omitted variable and in this case the direction of omitted variable bias for disposable income is toward zero, since disposable income would likely trend positively with the number of vacations. This is unfortunately bad news for our model; since given the omitted variable, our statistical significance would decrease. This lends to the idea that our effects are less likely to be real, i.e. they might just be entirely an artifact of omitted variable bias.

Another potential omitted variable of interest could be political leaning. As seen from the interpretation of the negative correlation between vaccinations and vacations taken, states which were not in favor of vaccinations were also not in favor of lockdowns or limiting their freedoms (e.g. vacationing). Due to the categorical nature of this omitted variable the omitted variable bias could go either way, i.e. liberalism being the omitted variable vs conservatism being the omitted variable. With liberalism as the omitted variable we get a negative correlation and thus the bias is away from zero, while with conservatism the correlation is positive and thus the bias is toward zero. Perhaps a better metric, which would be highly collinear with these two variables, would be belief in science. Greater belief in science would trend negatively with the number of vacations. This would produce an omitted variable bias away from zero benefitting the significance of our model.

Conclusion

In conclusion, we find that the models overall trended negatively overall towards vacation trips per capita in the U.S. and that our limited model, model 1, was our most significant model in showing such as it generated the lowest Adjusted R^2 of .234 and our highest Residual Standard Error (RSE) of .392.

Likely indicating that people are less inclined to travel about on vacations that are particularly far from their home state. This would also suggest that in spite of vaccinations being administered and millions of people being fully vaccinated, it is also revealing that people that are fully vaccinated are behoved to not travel as much, due to the large number of individuals that are still yet to be vaccinated. However, with the only significant indicator being Vaccination rate proportional to the given state's population, all other factors show to be largely insignificant to the idea that people will travel less on further vacations.

Our exploration into whether the state's performance against the COVID-19 virus and the individual state's actions to counteract the spread, thus affecting vacationing, gave mixed results. We were able to observe that even though vaccination_rate_prop was a negative affecting variable, it was indeed significant, generating a p-value of 0.0008. However, when we added the the additional variables of Cases_rate_prop , stay_at_home_prop, total_mask_mandate_duration_prop, Deaths_rate_prop, and Weekly_UI_Amount, we see that each of these variables were not significant to indicating the trend direction statistically as each of the p-values for these variables were well over .04 with the total_mask_mandate_prop nearing an even 1 at .96151. Also it is very likely that there causal relationships between these variables is the reason for them not being significant within this model 2, whether it being the potential political motivations in general of a given state leading to the lesser mask mandate or one being much longer and causing fewer people to be inclined to travel or disregard such warnings that have been presented before them or cases being much higher in other states due to the mask mandate not being in effect long enough for some states causing the deaths and cases to be higher or simply the population of a given state being the reason regardless of the mandate or any other policies enacted. Overall, with how low in significance the variables added for model 2, we do not gain more insight than our model 1 as to what the trend for travel proportions for the different states during this covid-19 crisis. Lastly, our model 3 adds in the remainder of the rates per day and totals per day as well as the eviction moratorium, stay at home length proportion, and mask mandate length proportion, we see that we introduce more variables largely insignificant to the regression model as our coefficients grow to be very large and do not indicate anything that wouldn't be better found from our model 1 or model 2, especially with issue of multicollinearity of the variables being present among the regression.