Lab 2: Vacations and Vaccinations, Summer 2021

Laura Chen, David Trinidad, Joe Villasenor

August 3rd, 2021

Contents

Introduction

With the COVID-19 pandemic continuing to disrupt everyday lives and habits, there have been extensive research studies on which aspects of our lives have been most impacted and what we can expect for the near future.

Under several restrictions in the state of California, communities were advised to remain cautious and vigilant, and to abide by shelter-in-place policies to mitigate the possibility of contact with the COVID-19 virus outside. [1] It has been quite some time since the initial rollout of the various vaccinations, along with the mutation of the new Delta variant which was first identified in December 2020. It has now become the predominant strain in the US. [2] In California, vaccinations and mask mandates, among other protective measures, have been generally encouraged in multiple counties. However, in June 2021, the state lockdown was terminated, and restrictions such as physical distancing, capacity limits on businesses, and the county reopening tier system were lifted. [3]

We are interested in investigating not policy, but vaccination and COVID testing rates which may have effects on community mobility levels, and what this could mean for travel and tourism in 2021. This analysis is not meant to be causal, as we are aware that there are a myriad of other factors that could potentially influence community mobility levels that are not within the scope of our current data.

The primary goal of this analysis is to identify several COVID-19 statistics and what effect they may have on community mobility levels. However, a limitation of our data is that we are investigating 58 counties in California, which may lead to our conclusions being slightly skewed. Additionally, different counties may have experienced COVID-19 at varying times and magnitudes, whether due to population, geographic, or socioeconomic factors.

The following section explains our main research question in more detail and formulates the initial hypotheses that we conduct our models against. We then provide context and transformations on the datasets used for this analysis and justifications for the changes. We will work through a limited model with one key variable of note and then include additional covariates to inform our modeling. After running our multiple regression models, we will evaluate which one is most effective, and present our findings and takeaways.

Research Question

Research Question: Given the marked changes that COVID has brought upon peoples' everyday lives, we are looking to investigate if there is a statistically significant relationship between vaccination rates , patients testing positive for COVID-19, patient deaths, testing and community mobility and travel tendencies between counties in California and what this could entail for tourism in summer 2021. We will be focusing on a specific point in time to avoid delving into what could turn into a time series model.

β_1: Percentage of Population with full dosis of vaccination β_2: COVID-19 related deaths β_3: COVID-19 total number of people tested β_4: Population testing positive for COVID-19 Y: community mobility levels

Our hypothesis is that there is a relationship between the above mentioned variables and community mobility levels. $Mobility = \beta_0 + \beta_1 * (CompletedVacc) + \beta_2 * log(Covid19deaths) + \beta_3 * log(Covid19tests) + \beta_4 * log(PositiveTests)$

A Model Building Process

Action Plan

The data source(s) we are working from is segmented into state-level/county-level appropriately. We will be cleaning and investigating the data in R and extracting the columns of note, as well as mutating new columns for needed variables. (Accounting Table)

Data Sources:

The COVID-19 Community Mobility Reports are updated daily by Google and include data procured anonymously through owners of Google Accounts that have Location History turned on. These reports are provided for public use and analysis and give insights into community mobility trends across multiple locations such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential areas. [COVID-19 Community Mobility Report] [Google Report] (https://www.google.com/covid19/mobility/)

The second dataset we leverage is the CDC COVID-19 Vaccinations by County, which is updated daily with information on vaccination rates and age groups of vaccinated individuals. The data is collected through all official vaccine partners, which includes but are not limited to retail pharmacies, long-term care facilities, and federal entity facilities. This provides pertinent insights at a more refined level on one of our variables of concern, vaccinations by county. CDC Data on Vaccinations

Lastly, the third dataset was pulled from the DSH California Covid-19 Patient Data, including multiple variables of interest for our model: total COVID-19 tests administered, total positive cases of COVID-19, and total deaths of patients who tested positive for COVID-19. The data is collected from DSH patients who receive treatment for COVID-19 at outside medical facilities, and data has been appropriately anonymized. DSH California Covid-19 Patient Data

What do we want to measure?

Description of the variables:

Mobility

Google's Mobility Report show movement trends by region, across different categories of places. foe each category in a region. If they didn't have enough data to confidently and anonymously estimate the change from the baseline, it was reported as a NA.

Baseline is defined as a normal value for that day of the week. It is the median value from the 5-week period from Jan 3 - Feb 6,2020 (Pre-COVID). The baseline isn't a single value -it's 7 individual values. The same number of visitors on 2 different days of the week.

Mobility is reported across: Retail & Recreation, Parks (Public garden, Castle, National forest, camp ground, observation deck), Transit Stations (Subway station, sea port, taxi stand, highway rest stop, car rental agency), Groceries and Pharmacies, Residential (Time spent at places of residence).

Series_Complete_Pop_Pct

Percent of people who are fully vaccinated (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where recipient lives

Cumulative_deaths

Number of cumulative deaths from first death reported in 2020 until day selected.

Cumulative_total_tests

Number of cumulative COVID19 diagnostic tests from first patient tested in 2020 until day selected.

Cumulative_positive_tests

Number of cumulative COVID-19 positive tests from first case reported in 2020 until day selected.

Table 1: Accounting Table

County	$\Delta Rtl/Rec$	$\Delta Parks$	$\Delta Transit$	CV	Deaths	Tests	+ Tests
Alameda County	-23	15	-57	64.2	1265	3086003	111949
Contra Costa County	-17	-2	-52	63.3	831	1834855	83768
El Dorado County	14	150	-42	49.4	116	213462	11151
Fresno County	-2	13	-9	42.3	1742	1393498	122963
Humboldt County	8	118	-50	50.5	51	162870	5598
Imperial County	-21	-65	-10	56.2	743	344683	33035
Kern County	3	10	-1	34.9	1353	1220697	112045
Kings County	-5	1	-38	29.5	247	407223	27183
Lake County	13	68	-31	43.5	65	80579	4893
Los Angeles County	-18	-12	-36	53.1	24683	24126420	1595755
Madera County	2	43	94	37.1	244	291256	19224
Marin County	-8	24	-54	73.1	239	580224	16246
Mendocino County	23	117	-20	51.3	49	112536	4870
Merced County	0	29	-8	31.1	471	413210	34591
Monterey County	-9	20	-11	51.2	524	677237	49163
Napa County	-14	76	-36	61.1	81	298701	11109
Nevada County	10	77	-29	48.7	72	120938	5461
Orange County	-15	-2	-39	55.2	5144	4282143	320184
Placer County	-5	51	1	51.2	306	480543	26391
Riverside County	-8	-30	-39	41.6	4520	3289798	355871
Sacramento County	-14	20	-26	49.1	1718	2218663	130230
San Bernardino County	-8	-18	-20	39.8	5197	3317711	353678
San Diego County	-12	8	-28	43.4	3798	5299628	324938
San Francisco County	-37	-13	-63	68.8	562	2312213	51244
San Joaquin County	-2	38	-10	39.0	1442	1149309	87455
San Luis Obispo County	8	41	-34	46.5	260	579157	23819
San Mateo County	-20	23	-45	67.0	540	1914060	56011
Santa Barbara County	-10	14	-36	53.6	466	709649	40982
Santa Clara County	-24	14	-56	68.4	2089	4639997	140953
Santa Cruz County	-10	53	-59	60.6	207	479863	17400
Shasta County	4	100	-33	35.2	212	231798	10264
Solano County	-6	-9	-40	48.7	260	782972	37178
Sonoma County	-15	33	-48	61.2	329	800916	35684
Stanislaus County	-3	36	-11	35.4	1032	744909	66640
Tulare County	-9	43	-6	35.2	854	636441	57290
Ventura County	-15	-12	-46	55.1	1036	1604607	101226
Yolo County	-21	14	-25	54.0	215	576261	15074
Notes							
Rec: Recreation							
Rtl: Retail							
CV: Completed Vaccination							

Exploratory Data Analysis

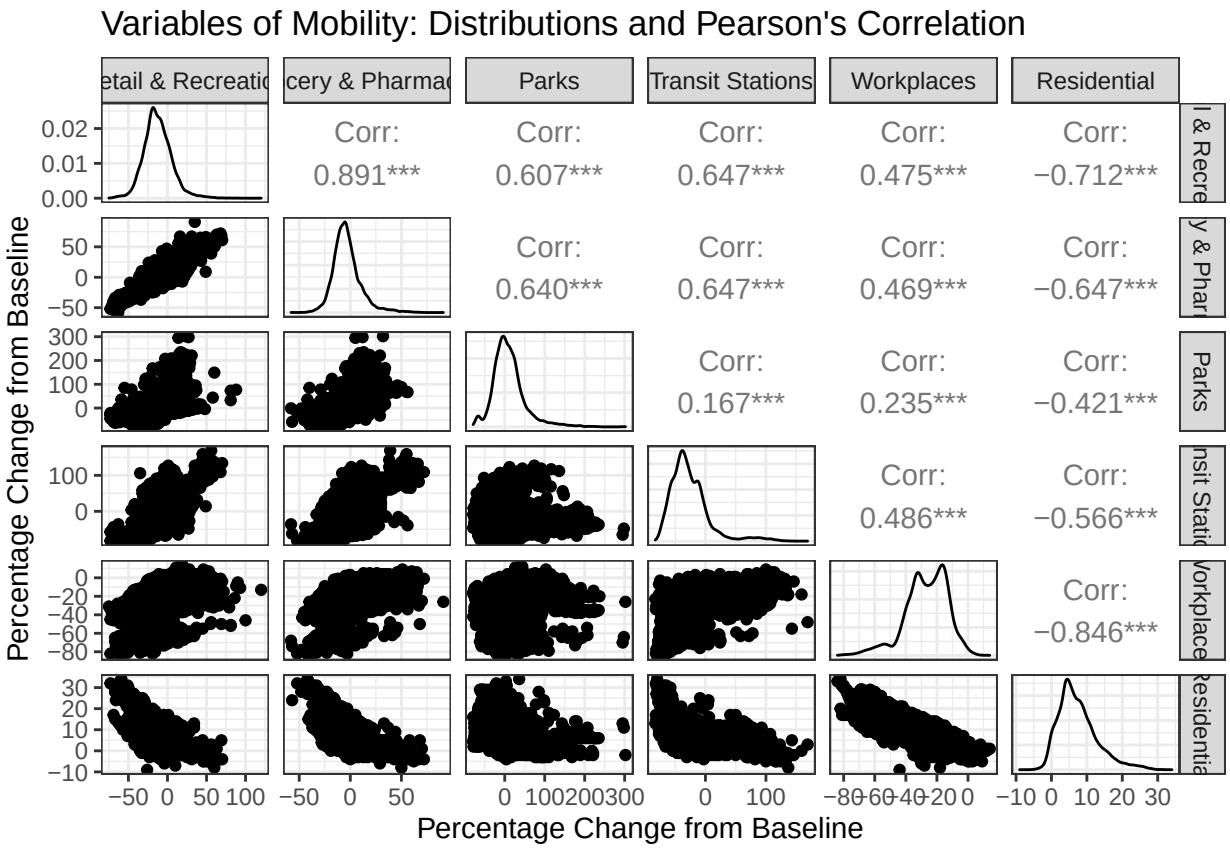
How to define our dependent variable? Mean Mobility

Mobility is measured, as explain in the introduction, as a percentage change from the baseline. The baseline is considered a 5 month's average, pre COVID. We will summarize five variables in one: "Mean Mobility"

Covariates and collinearity (Pearson Correlation) for Dependant Variable

Question we want to answer: "Is Retail & Recreation (primary related to vacations) correlated with Parks and Transit Stations mobility?" If the answer is yes, then we can bundle them in one single mobility metric, if not, then we will exclude them since we are interested in measuring "Vacations", not the effect in mobility due to shelter-in-place or mobility in transit stations due to changes in non-vacation mobility.

Variables Correlation



Retail correlates mildly positive with Parks and Transit Stations. Transit Stations does not correlate with Parks. Therefore it is safe to assume that we can bundle these three in one single metric.

1. Most of Retail and recreation is due to Grocery and Pharmacy mobility. Not really a "Recreation"
2. Retail and Recreation is negatively correlated to Residential, so we will exclude it.
3. Retail and Recreation is mildly related to Parks and Transit Stations, so we will keep them.
4. Retail and Recreation is lowly correlated to workplaces so we will also exclude it.

In conclusion, or our mobility related to vacations, we will only use Retail & Recreation (Although we know we have some room for error because this is mostly related to groceries and pharmacies movement), Parks and Transit Stations.

Transformations

Table 2: Accounting Table

County	$\Delta Retail/Recreation$	$\Delta Parks$	$\Delta Transit$	$\mu mobility$
Alameda County	-23	15	-57	-21.67
Contra Costa County	-17	-2	-52	-23.67
El Dorado County	14	150	-42	40.67
Fresno County	-2	13	-9	0.67
Humboldt County	8	118	-50	25.33
Imperial County	-21	-65	-10	-32.00
Kern County	3	10	-1	4.00
Kings County	-5	1	-38	-14.00
Lake County	13	68	-31	16.67
Los Angeles County	-18	-12	-36	-22.00
Madera County	2	43	94	46.33
Marin County	-8	24	-54	-12.67
Mendocino County	23	117	-20	40.00
Merced County	0	29	-8	7.00
Monterey County	-9	20	-11	0.00
Napa County	-14	76	-36	8.67
Nevada County	10	77	-29	19.33
Orange County	-15	-2	-39	-18.67
Placer County	-5	51	1	15.67
Riverside County	-8	-30	-39	-25.67
Sacramento County	-14	20	-26	-6.67
San Bernardino County	-8	-18	-20	-15.33
San Diego County	-12	8	-28	-10.67
San Francisco County	-37	-13	-63	-37.67
San Joaquin County	-2	38	-10	8.67
San Luis Obispo County	8	41	-34	5.00
San Mateo County	-20	23	-45	-14.00
Santa Barbara County	-10	14	-36	-10.67
Santa Clara County	-24	14	-56	-22.00
Santa Cruz County	-10	53	-59	-5.33
Shasta County	4	100	-33	23.67
Solano County	-6	-9	-40	-18.33
Sonoma County	-15	33	-48	-10.00
Stanislaus County	-3	36	-11	7.33
Tulare County	-9	43	-6	9.33
Ventura County	-15	-12	-46	-24.33
Yolo County	-21	14	-25	-10.67

Independent Variables (Exploratory Data Analysis)

- What is our control? Our short model will be our control. $Mobility = \beta_0 + \beta_1 * (CompletedVaccinations)$

Omitted Variables

Officially Reported Cases, Deaths and Tests [CA.gov dataset for COVID 19 cases]: These are official cases reported by only healthcare institutions and captured in the CA government data as opposed to the non reported (Cumulative Cases, Cumulative Deaths, and Cumulative Tests), which covers both official cases and non-official cases. Because the sample size is larger, our team used the non-reported features for our analysis

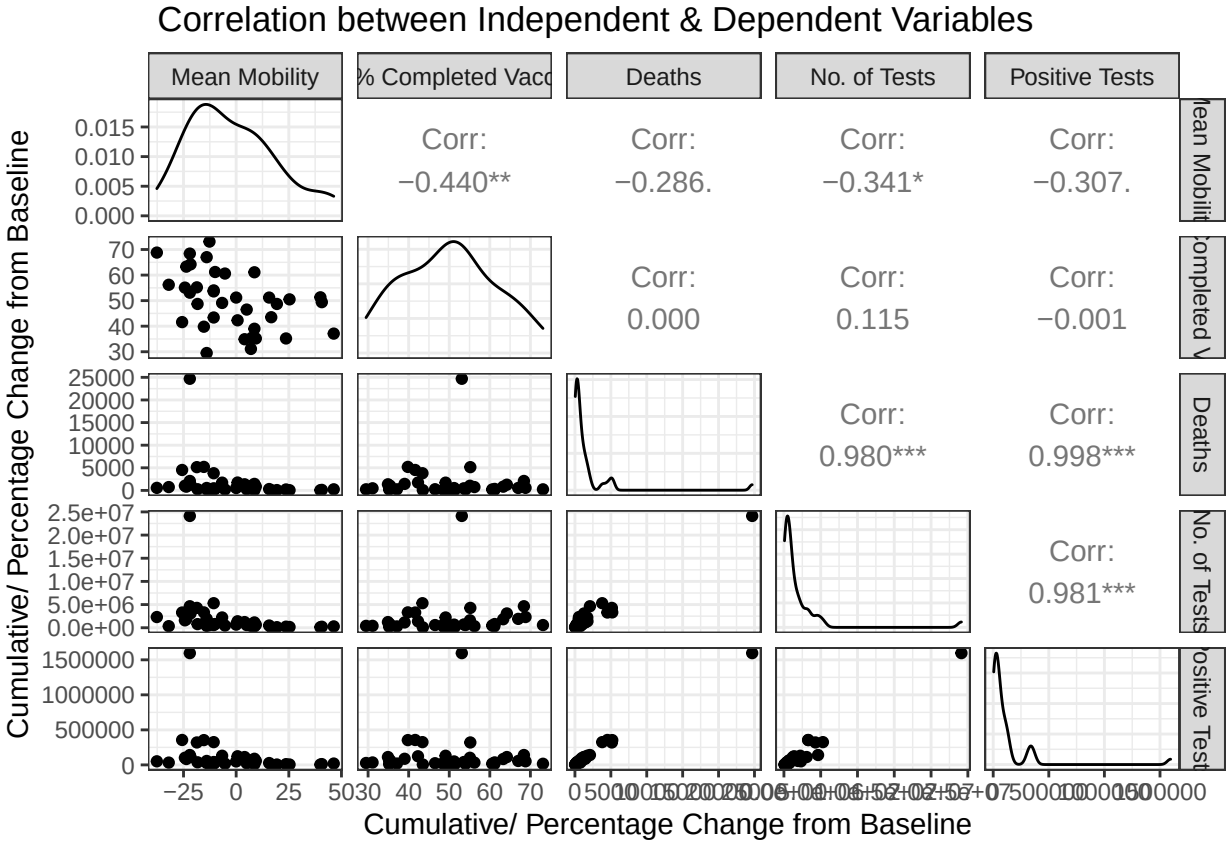
for the X2, X3, and X4 variables. Since the state manages the data, we assumed any risk to validity or bias in the data feature to be minimal against our results.

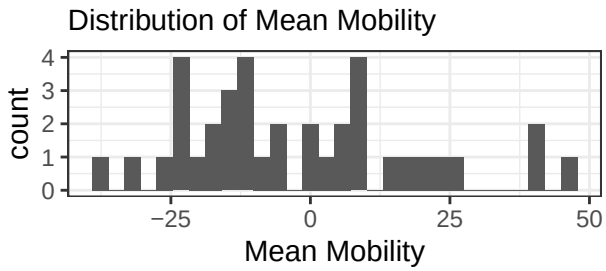
Age-Specific Features [CDC vaccination dataset] - For our independent X1 variable our team omitted features that are age-specific such as "Series_Complete_12PlusPop" (completed vaccination series percentage for ages 12+ population). At this stage, we feel that age granularity at this stage of our analysis is not necessary. Single Dose Vaccinations [CDC vaccination dataset]: Single dose vaccination features such as "Administered_Dose1_Recip" were left out of our analysis based on the assumption that patients are less likely to travel at the risk of missing the window for their second shot. Doctors continue to stress the importance of not missing the 2nd shot window to ensure efficacy against the Delta Variant and due to the limited supply of the vaccine.

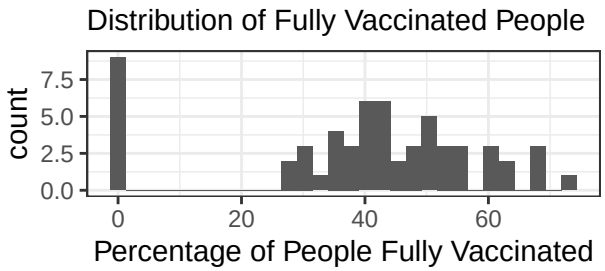
- How dropping out counties might be random or systematic? We dropped a total of 12 counties that didn't have enough data to comply with our rule for mean mobility. This process was not completely random, since it is obvious that more rural counties would've had less access to the technology needed to report certain type of mobility. This, although not ideal, was the way we defined the experiment and trying to change dates, variables to avoid dropping counties to fix it to dates that had more data, would've not been a good approach.

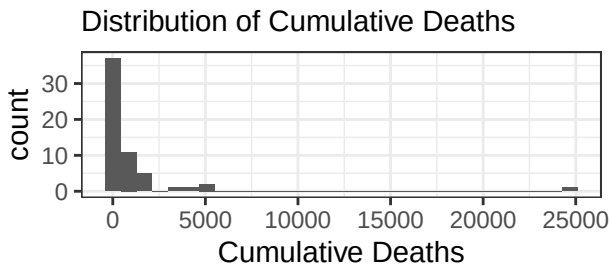
Exploratory Data Analysis Dependent and Independend Variables

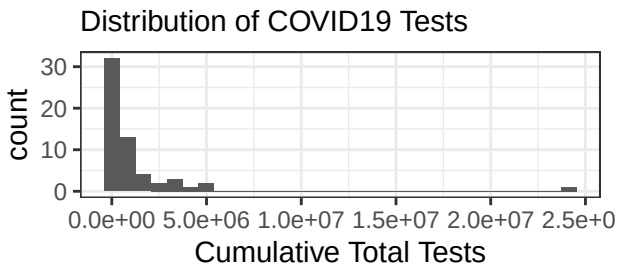
We want to see the correlation of Mean Mobility with the independent variables we have chosen for our analysis. In this section we will look at their distributions, correlations and needed transformations for a linear regression model.

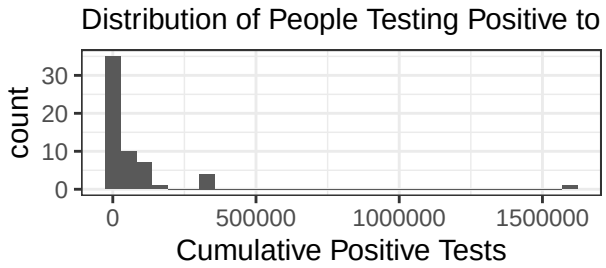












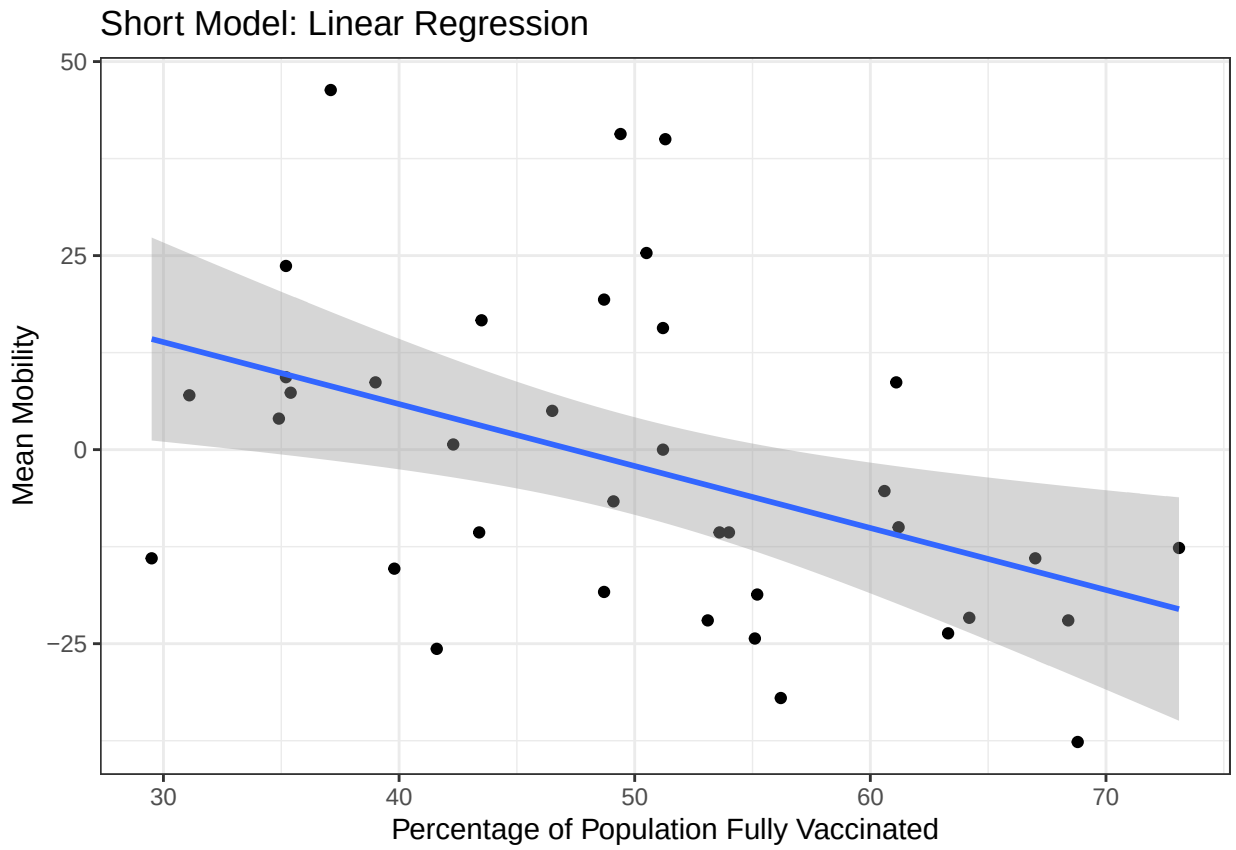
Insights after Initial EDA

1. Only Mean Mobility and Percentage of Population Vaccinated have normal distributions
2. Cumulative deaths, total tests and positive tests have skewed distributions to the left.
3. There is a point that seems to be an outlier but it is not, it is the county of LA that because of it's population and different demographics had the highest incidence of deaths, testing and positive tests. We cannot eliminate this point to normalize the distribution.
4. There is no strong correlation between mobility and any of our independent variables. Elder populations and younger might be skewing the data, since the first had reduce mobility due to higher risk and the younger, besides not having vaccinations until later, they were lower risk of death and complications.

Short Model Linear Regression

There seems to be a negative linear correlation between mobility and percentage of people fully vaccinated, therefore our baseline and short model will be the following:

$Mobility = \beta_0 + \beta_1 * (CompletedVaccinations)$



Insights of the short model

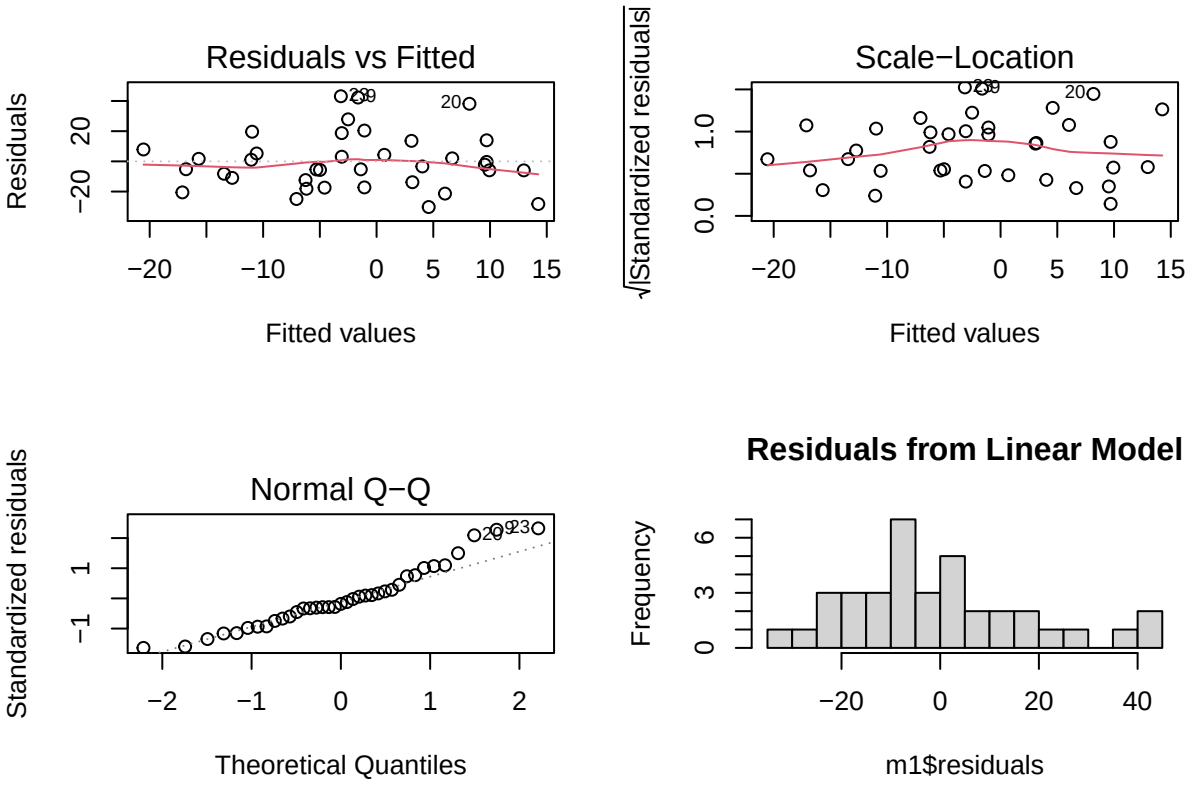
We expected that mobility would be higher if people had their vaccination scheme complete, it was the opposite. The lower the percentage of the population with a full scheme of vaccination complete, the more mobility they had! Maybe what we explained above in our hypothesis of younger and older groups skewing the data in opposite directions.

This would've been because they never cared for their vaccination since the beginning or because some counties experienced more mobility due to flex working "work from home" and moving to more rural counties?

On top of that if we do a linear model of these two variables we can say that a decrease in -0.798 in percentage of population vaccinated, we will get one more point in mobility overall, or in other words with every 0.798 percentage of population vaccinated, we will have a change of mobility of 1 in the mean of all services.

If no-one had their vaccination scheme complete, we would assume a mean mobility percentage of 37%

Statistical Analysis of the Short Model



1. Residuals vs Fitter seems to be linear, this is important for a regression model since the residuals are estimates of the error of the estimates in our model.
2. Normal Q-Q plot helps us to assess if the residuals are normally distributed. Fitted values have a little quadratic behavior, this because in the right tail of distribution we have LA County for which the model is trying to adapt to.
3. Scale-Location vs root sqr of Standardized residuals helps us understand if the errors are normally distributed, which we can see they do have some curve in both tails, but could be considered minimal.
4. We can also see in the last graph the residuals from the linear model in an histogram. Ideally we would like it to be close to zero, so that it is closer to the BLP, however in this case we see some residuals.

Let's analyze even further what this means in the following section.

```
##
## ================================================
##                             Dependent variable:
##                         ---------------------------
##                                mean_mobility
## ------------------------------------------------
## Series_Complete_Pop_Pct          -0.798***
##                                   (0.245)
##
## Constant                         37.799***
##                                  (13.407)
##
## ------------------------------------------------
## Observations                        37
## R2                                 0.193
## Adjusted R2                        0.170
## Residual Std. Error          18.838 (df = 35)
## F Statistic                8.383*** (df = 1; 35)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

We can see the model can reject the null hypothesis, which is that there is no correlation between our dependent variable (mean mobility) and our independent variable (Percentage of Population with Completed Vaccinations), since our p value is less than 0.01 (F Statistic) On top of that we can also see in our R2 value that the variation of a dependent variable is explained by the independent variable is pretty low, therefore we would need to expand on the model to improve it, since this single variable alone, is not doing a great job to do so. Let's do it in our extended models.

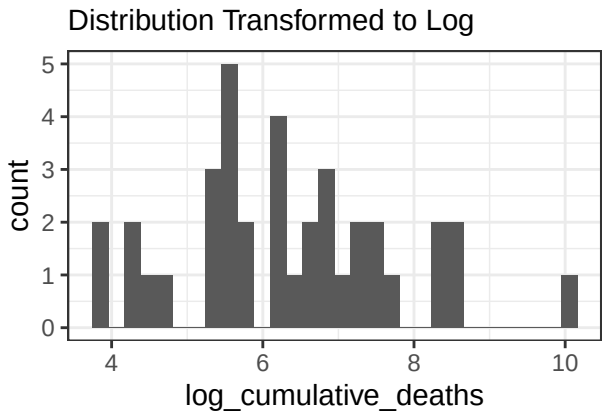
Extended Models

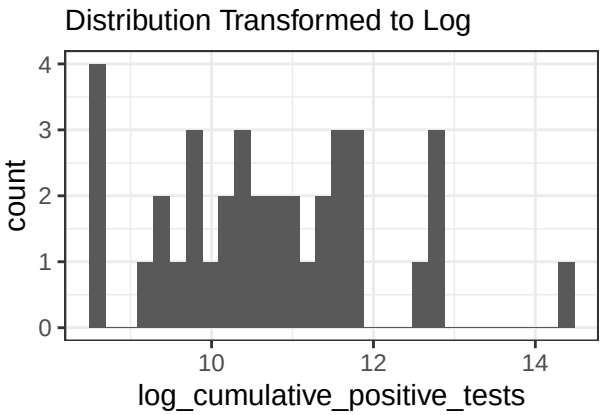
We add more models to include key explanatory variables and covariates in an effort to advance our modeling efforts and analyze if it's introducing other issues like confounding.

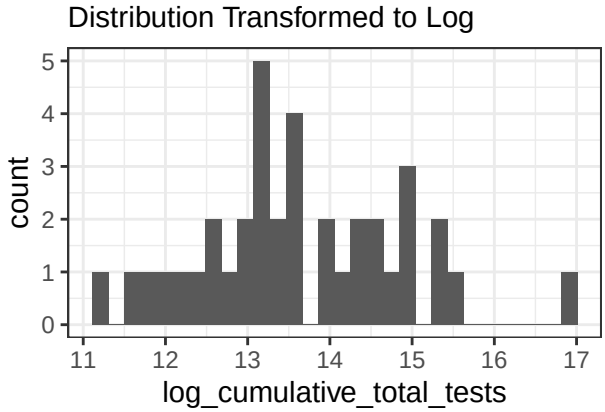
Our model is described below:

note: the highly skewed variables have been modified to a logarithmic scale in order to correct for the skewness mentioned above.

$Mobility = \beta_0 + \beta_1 * (CompletedVacc) + \beta_2 * log(Covid19deaths) + \beta_3 * log(Covid19tests) + \beta_4 * log(PositiveTests)$







In the distributions transformed we can see that the logarithmic transformation helped to get rid of the skewness and behave more in a normal distribution

Regression Table

Note: we are using robust standard errors in all our calculations.

```
##
## ===================================================
##                          Dependent variable:
##                 ------------------------------------
##                              mean_mobility
##                    (1)       (2)       (3)       (4)
## ---------------------------------------------------
## %Vacc Completed -0.798*** -0.863*** -0.722**  -0.814**
##                  (0.276)   (0.203)   (0.292)   (0.329)
## Log Deaths                -8.991***  -5.171    0.0002
##                            (1.623)   (5.859)  (10.173)
## Log Test                             -4.716    -0.245
##                                      (6.945)  (10.017)
## Log + Test                                     -9.393
##                                               (15.036)
## Constant        37.799**  97.896*** 130.998** 142.654**
##                 (14.124)  (15.018)  (51.046)  (54.799)
## ---------------------------------------------------
## Observations       37        37        37        37
## R2                0.193     0.576     0.582     0.587
## Adjusted R2       0.170     0.551     0.544     0.535
## ===================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

In the table above we see that we have significant values in model 1 and model 2. It seems that model 3 and 4 have added not significant increase in accuracy but other issues like covariability that we will analyze below.

1. The big difference between model 1 and 2 is that model's 2 R-squared is better. This indicate the percentage of the variance in the dependent variable that the independent variables explain collectively. Therefore model 2 seems more appropriate.
2. Model 2 reduces the residual standard error, therefore reduces the difference between observations and the predicted values.
3. All models have a F Statistic that is significant ($p<0.01$) which can help us reject our null hypothesis. Null hypothesis being that all the coefficients in the model are equal to zero, in other words, none of the predictor variables have significant relationship with the response variable.

In conclusion we will move forward to analyze further model 2, that relates mobility with full scheme of population vaccinated and cumulative number of deaths due to COVID-19.

IID

We do run into some problems with the data being independent, as we are investigating counties in California, and there will be some relationships by geographic location between them. Since we chose to measure mobility levels by averaging mobility levels to different destinations, neighboring counties could likely have similar levels of mobility driven by naturally occurring conditions, similar infrastructure in transportation, or even political affiliation of the governing members. While there are many counties in California with different demographics, we can see that counties with similar socioeconomic status and population demographics like race and age tend to be located next to one another. A county being rural, urban, or suburban would also be a factor that introduces dependency to our data. Hence, similar counties could be grouped to mitigate the effect of the dependence in future models.

No Perfect Colinearity (Variance Inflation Factor)

Perfect collinearity is defined as having at least one variable that another can explain through a linear relationship, which would lead to a perfect correlation between the two variables. The OLS estimator cannot be estimated if we have ideal collinearity, as the coefficient of the first variable will capture the effects of both variables rather than just a single effect. Additionally, with solid collinearity, our standard errors would become infinitely larger, where our sample means would be spread widely around our population mean. Thus our sample would not be wholly representative of the population. We evaluate our model 2 by calculating the Variance Inflation Factor and seeking to observe values lower than 4, implying minimal collinearity. As evidenced by the VIF, our model does satisfy the assumption of no perfect collinearity.

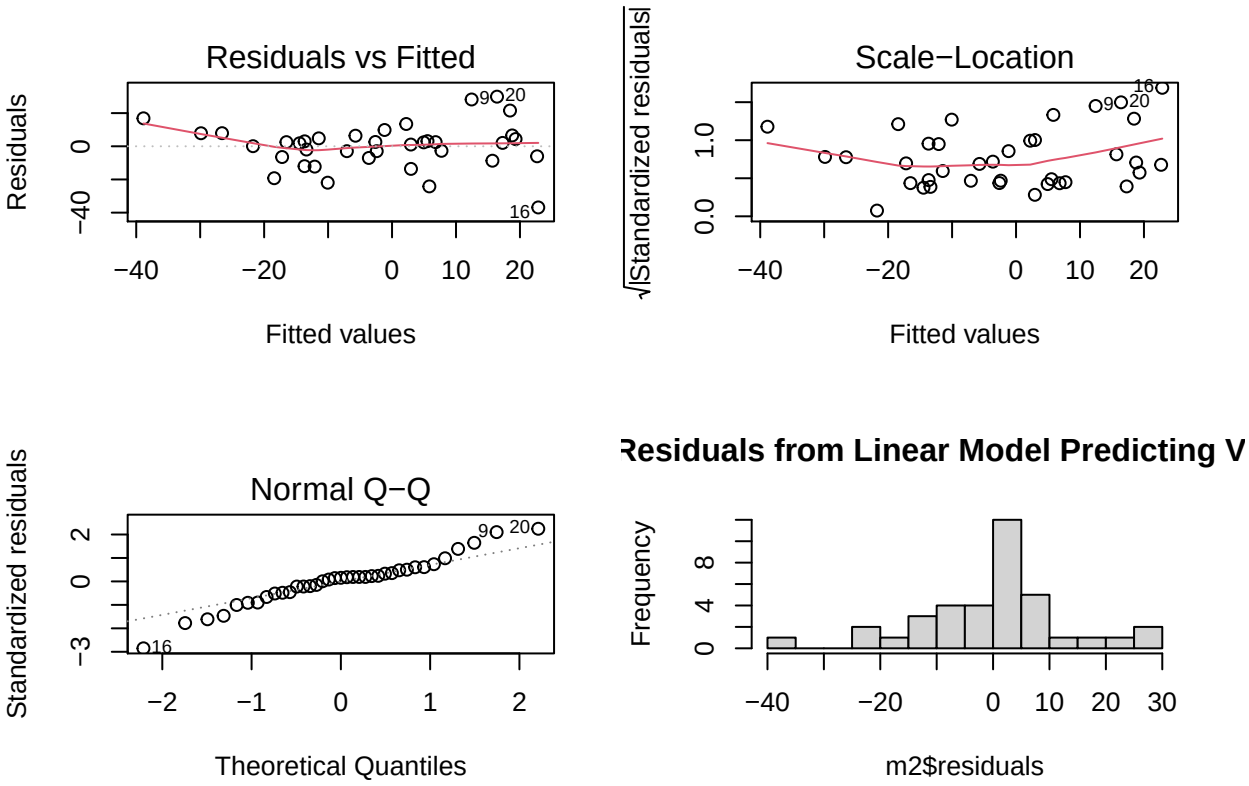
```
##
## Variance Inflation Factor
## ============================================
## Series_Complete_Pop_Pct log_cumulative_deaths
## --------------------------------------------
## 1.003                   1.003
## --------------------------------------------
```

```
##
## Variance Inflation Factor >4
## ============================================
## Series_Complete_Pop_Pct log_cumulative_deaths
## --------------------------------------------
## FALSE                   FALSE
## --------------------------------------------
```

As expected the results are negative.

Linear Conditional Expectation, Homoskedastic Errors and Normally Distributed Errors

Another assumption for a CLM is met if the explanatory variables have a linear conditional expectation. Linear conditional expectation can be validated by checking our Residuals vs. Fitted plots of the model for each explanatory variable. We can see that the fitted line is more or less linear. At higher values, our model does fulfill linearity quite nicely. Still, we have an issue on the left side of the line, which may be due to the transformation of the zero deaths to a logarithmic where we added a one to transform the zero values, leading to a small spike at a lower value than at gradually converges. Should the above issue be addressed by re-evaluating and potentially re-transforming the variable, we could satisfy the Linear Conditional Expectation more strongly.



1. Residuals vs Fitted: Make a good linear approximation, we have a little issue in the left side maybe due to the transformation of the zero deaths to a logarithmic that we added a one to transform the zeros.

2. Normal Q-Q: there's a little bit of a bilateral end issue both on the right and left, we will further quantify this in the next section.

3. Scale-Location: shows if the residuals are spread equally along the ranges of predictors, this means equal variance or homoscedasticity. We are looking for a horizontal line, we still see some effect in the tails so we will need to further analyze.

Homoskedasticity is defined as constant variance among residuals in the regression model, where even as the explanatory variable changes, the error sees little to no variance. We test if our model satisfies the homoskedasticity assumption by observing the Scale-Location plot of the model. To satisfy homoskedasticity, the plotted line of the Scale-Location plot should be horizontal, but as we can observe, our model has taken on a slightly parabolic shape. Additionally, the spread around the red line should be randomly dispersed with no clear pattern across all points. We can run a Breusch-Pagan test as well to confirm if our model meets the assumption of homoskedasticity.

```
##
##  studentized Breusch-Pagan test
##
## data:  m2
## BP = 3.8543, df = 2, p-value = 0.1456
```

As we can see, the p-value is 0.1456 which means we fail to reject the null hypothesis of homoskedasticity, so the variance among residuals is evenly spread. If the variance was not evenly distributed among residuals, then it would imply that the standard errors vary across values of the explanatory variables, leading to unreliability of the model.

Shapiro Test and Variance Co-Variance of Heteroskedastic Errors

The last assumption of a CLM is having normally distributed errors. This assumption can be validated through a visual inspection of the Normal Q-Q plot. As we can see, there are bilateral end issues on both ends, so we also run a Shapiro-Wilk normality test on the residuals to investigate further. We run the Shapiro-Wilk normality test on the residuals and come out with a p-value of 0.2589, failing to reject the null hypothesis, which means that the distribution of the residuals may be approximately normal. Thus, we do satisfy the assumption that our errors are normally distributed with a mean of zero. If they were not, then the difference between our model and population would not be close to zero, meaning that it is not an accurate representation of the population.

Shapiro Test:To tell if a random sample came from a normal distribution and Variance Co-Variance :calculation of robust standard errors Heteroskedastic

Model 1:

```
##                          (Intercept) Series_Complete_Pop_Pct
## (Intercept)               179.744069             -3.20136877
## Series_Complete_Pop_Pct    -3.201369              0.06015206
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.9494, p-value = 0.09228
```

```
##
##  Box-Ljung test
##
## data:  m1$residuals
## X-squared = 19.892, df = 5, p-value = 0.001309
```

Model 2:

```
##                          (Intercept) Series_Complete_Pop_Pct
## (Intercept)               445.708021             -4.37181133
## Series_Complete_Pop_Pct    -4.371811              0.05956257
## log_cumulative_deaths     -31.885819              0.18422973
##                          log_cumulative_deaths
## (Intercept)                          -31.8858187
## Series_Complete_Pop_Pct                0.1842297
## log_cumulative_deaths                  3.3783334
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.96343, p-value = 0.2589
```

```
##
##  Box-Ljung test
##
## data:  m2$residuals
## X-squared = 12.642, df = 5, p-value = 0.02698
```

ANOVA

Analysis of Variance The null hypothesis (H0) of the ANOVA is no difference in means, and the alternate hypothesis (Ha) is that the means are different from one another. One-Way-Anova: Mean Mobility ~

Completed Vaccination

```
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## Series_Complete_Pop_Pct   1   2975  2974.9   8.383 0.00648 **
## Residuals                35  12421   354.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way-Anova: Mean Mobility ~ Completed Vaccination + Cummulative Deaths

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## Series_Complete_Pop_Pct   1   2975  2974.9   9.059 0.0049 **
## cumulative_deaths         1   1255  1255.1   3.822 0.0589 .
## Residuals                34  11166   328.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Second model has a slighter better performance when we compare the analysis of variance.

Conclusion

The overarching goal is to determine whether or not there is a statistically significant relationship between vaccination rates (within CA counties) and average mobility rates (travel between counties) using the available data sources managed by the CA state and federal government. To do this, our team identified key COVID-19 statistics and measured the statistical relationship mean mobility. Health authorities such as the FDA and CDC strongly suggest individuals continue to shelter in place and postpone all travel until fully vaccinated. [2] This includes vacation/leisurely travel, the use of public transportation, as well as other activities that may run the risk of contracting or spreading the disease.

Our null hypothesis (Ho) assumes no relationship between mean mobility rates (Y) and our independent variables X1, X2, X3, X4. According to our results in our regression table, All models have a significant F Statistic ($p<0.01$) which enables us to reject our null hypothesis.

We observe a statistically significant relationship between our dependent variable (mean mobility) and our independent variables (cumulative vaccination, cumulative tests, positive COVID test results, and COVID-related deaths). However, between our 4 models, model 2 provided our team with the best estimation for the overall mobility.

Key Learnings

Based on our analysis, we learned that there is a strong collinearity between COVID death rates and mean mobility. As we continue to add more data to our 2nd model, we can predict that mean mobility rates will decrease as COVID-related deaths increase. Further analysis might have to be done to understand the effects of the age skew that might be present in the models: younger people not getting vaccinated at the same time as the older population and the latter being more prone to complications of COVID, therefore, being more cautious about going out. The younger population not being inclined to severe complications might have skewed the data in mobility, so we think a further analysis with age as a variable is of essential value to this research.

Recommendations

Analyze the younger adult population. According to a report from CNN, individuals between ages 30 to 39 were more likely than average to miss their second dose of the vaccine. Meanwhile, children under 18 were least likely to skip their second dose. [4] This variable may have skewed our data because many individuals choose not to complete the dosing series either in fear of the common symptoms after the second shot or those who feel that a double dose is necessary.

Analyze the individual features of the mobility data. For simplicity of measuring establishing our Y variable (mobility), our group combined the baseline changes for all features. In other words, we treated the percent change from baseline at retail recreations, grocery, and pharmacy, parks transit stations, residential, and workplace as the same. This generalization of the data features may likely have led to our model's inaccurate results. To further validate our measurement for the mean mobility, we would need to observe how each feature will affect the output prediction of our four models. Analyze whether or climate data. Changes in weather or climate patterns might influence mobility. For example, individuals are less likely to leave their homes during rainy or snowy weather. To mitigate this bias from our model, the next step would be to incorporate weather conditions as part of our sampling model.

References

[1] Procter, Richard. "Remember When? Timeline Marks Key Events in California's Year-Long PANDEMIC Grind." CalMatters, 4 Mar. 2021, calmatters.org/health/coronavirus/2021/03/timeline-california-pandemic-year-key-points/.

[2] Katella, Kathy. "5 Things to Know about the Delta Variant." Yale Medicine, Yale Medicine, 3 Aug. 2021, www.yalemedicine.org/news/5-things-to-know-delta-variant-covid.

[3] California, State of. "Safely Reopening CALIFORNIA." Coronavirus COVID-19 Response, 30 July 2021, covid19.ca.gov/safely-reopening/.

[4] McPhillips, (2021) "More than 1 in 10 people have missed their second dose of Covid-19 vaccine" CNN https://www.cnn.com/2021/06/24/health/missed-second-doses-delta/index.html

[5] Aragon, (2021), California Dept of Health, "Travel Advisory" California Department of Public Health, https://www.cdph.ca.gov/programs/CID/DCDC/pages/COVID-19/Travel-Advisory.aspx

[6] Lovelace, CNBC, (2021)" CDC reverses indoor mask policy, saying fully vaccinated people and kids should wear them indoors" https://www.cnbc.com/2021/07/27/cdc-to-reverse-indoor-mask-policy-to-recommend-them-for-fully-vaccinated-people-in-covid-hot-spots.html