DATA 102 Final Project Written Report

Author: Jing Rong, Eddie Guo, JP Thrasher, Grey Xu

1. Data Overview

1.1. Google Community Mobility data

The data shows how visitors to (or time spent in) categorized places change compared to our baseline days, and it records the movement or mobility across different categories. A baseline day represents a *normal* value for that day of the week. The baseline day is the median value from the 5 week period Jan 3 – Feb 6, 2020. This data could be referred to as a census data set because it collects the anonymized sets of data from users across the states who have turned on the Location History setting to present the whole population.

The granularity of our Google mobility data is based on the intervals of dates, and each row represents the mobility percentage of change from the baseline on a particular date and place. Such a daily value-based model allows us to obtain enough data to implement causal inference or linear regression model and thus help to answer our research questions.

There is no personally identifiable information in this dataset, such as an individual's location, contacts or movement will be released to the public. People who have Location History turned on can choose to turn it off at any time from their Google Account. The data collector in Google also uses the same world-class anonymization technology to keep the activity data private and secure, which includes differential privacy by adding artificial noise to the datasets and thus prevent the identification of any individual person.

However, there might be selection bias because all of the google users are required to give the company the permission to collect their personal location information. Thus, other mobile users who are not willing to report their information could be more leaning towards a certain age group or working positions that might be confused with mobility. There may be some measurement error because some of the counties have missing values and may cause our result to be biased. There isn't any problem with the convenience sampling as it aims to capture the information from the whole population with abundant relevant data to ensure such an issue would not happen.

1.2. USEPA Air quality data

The second data set is from USEPA air quality monitored data. Ambient (outdoor) concentrations of pollutants are measured at more than 4000 monitoring stations owned and operated mainly by state environmental agencies. The agencies send hourly or daily measurements of pollutant concentrations to EPA's database called AQS (Air Quality System). This data set should be census data because it attempts to gather information about every individual county from its monitored air quality stations.

For the US EPA Air Quality Data, it allows us to display and download monitored hourly, daily, and annual concentration data, AQI data, and speciated particulate pollution data. Specifically, I chose to download the daily level data on the NO2 concentration level. On the main page of interaction, we could select the queries daily air quality summary statistics for the criteria pollutants by monitor. We can get data for specific monitors or all monitors in a city, county, or state by selecting the targeting study places. The granularity of US EPA NO2 concentration level data is also based on the time when a particular monitoring was processed. Each row represents the monitoring value of NO2 concentration level given a particular place and time. Since there are six measurements for each day and for each place, we take the average concentration for all these measurements and transform it to a single value for each day. This helps us to match the independent variable and the dependent variable into the same scale.

For the US EPA air quality data, none of the three concerns (selection mibas, measurement error, convenience samples) is relevant to our data as it is a national wide air quality data collected by the government. Also, all of the monitor stations have been implemented with sensitive machines to maximize the measurement accuracy.

1.3. COVID cases, vaccination, and policy

We also include three additional datasets on covid cases, vaccination, and policy for another research question. These dataset will be used in our first research question since we use covid policy as an instrumental variable when conducting causal inference between covid prevalence and social mobility.

COVID-19 Time-Series Metrics by County and State: This data is from California Open Data Portal. It shows covid cases, deaths, and tests by counties. The data should be considered as census because it records all covid cases, deaths, and tests in the population, which is a complete enumeration survey method. For the COVID-19 Time-Series Metrics by County and State data and COVID-19 Vaccine Progress Dashboard Data, they are directly downloaded from California Open Data Portal as csv files. The USA State Level COVID-19 Policy Responses data is downloaded from the public github repository of the research group in Blavatnik School of Government.

COVID-19 Vaccine Progress Dashboard Data: This data is from California Open Data Portal. In the dataset, This dataset summarizes vaccination data at the county level by county of residence. This dataset is also considered a census since it records all vaccines that are being distributed. For the COVID-19 Vaccination Progress dataset, the collection method implemented was not one that affected any one individual. Data was collected on the number of vaccinations shipped, doses delivered, and doses on hand for each facility where data was collected. Total vaccine progress was then calculated using these numbers which suggests that there was no issue with participants during the collection of this data.

USA State Level COVID-19 Policy Responses: This dataset is from research in Blavatnik School of Government. This dataset contains systematic information on several different common policy responses governments have taken during the pandemic. These policies are recorded on a scale to reflect the extent of government action and the scores are aggregated into a suite of policy indices. The dataset contains 21 indicators and a miscellaneous notes field organized into five groups: C (containment and closure policies), E (economic policies), H (health system policies), V (vaccination policies), M (miscellaneous policies). This dataset is also a census data since it records covid policies in all states in the US.

For the three covid related data, the above three concerns (selection mibas, measurement error, convenience samples) are also not relevant in the context of our data since the data are collected nationwide by the authority.

2. Research Questions

Question 1: What impact does the COVID-19 pandemic prevalence have on human mobility in public transportation in California?

By answering this question, we can gain insight into how our community is moving around differently due to COVID-19. Moreover, this would help public health officials identify which place categories covid spread most and send covid risk warning to the public. Causal inference is a good fit for answering Q1 since we are investigating causal relationship between covid prevalence and mobility and the two variables are empirically correlated with one another. We also use covid policy as an instrumental variable to isolate the confounding variables.

Question 2: How has human mobility in public transportation affected the NO2 concentration in Alameda County, San Diego County, Orange County, Fresno, County, Sacramento County, and Los Angeles County in 2020?

Answering this question helps to understand the mechanism of correlation between NO2 concentrations and human mobility in public transportation across different counties. This could help the California Environmental policy maker to decide which one of the six listed counties could be most beneficial when a certain regulation aims to reduce usage of transportation. Specifically, we want to use multiple hypothesis testing on 5 different counties to test mobility and NO2 relation, but we realize multiple hypothesis testing better comes with binary decision making to simplify our research questions. This could be a good fit for the problem if we set up a threshold for NO2 concentration and make it a binary outcome. Because multiple hypothesis testing allows us to control false discovery rate and family-wise error when we aim to test whether there is significant correlation between the mobility and NO2 concentration in each individual county.

3. EDA

3.1. List of EDA variables:

- Quantitative variables:
  - Daily Max 1-hour NO2 Concentration
  - Transit_stations_percent_change_from_baseline (mobility)
  - Covid and Vaccination Status in Alameda County

- Categorical variables:
    - County name
    - Covid Policies

3.2 Explanation of EDA

First, we visualized the data availability and the spatial visualization of the NO2 pollution across different counties in California. We find that there are several of the counties worth studying due to their high concentration level. These six counties are Los Angeles, Alameda, San Francisco, Fresno, Orange and Sacramento county. Also these six counties have a relatively abundant amount of data, which allow us to implement subsequent modeling procedures. From the mobility line plot, we noticed that there was a significant drop around March and April for all these selected counties. From the NO2 line plot, we could also see that the NO2 concentration decreased around these months, but it increased in a large amount in the later of the year while mobility only increased in a small amount. We thought that there may be some other confounding factors that are causing the NO2 concentration to increase. We may want to check out what other confounding variables may have higher correlation with NO2 concentration and include them into our model.

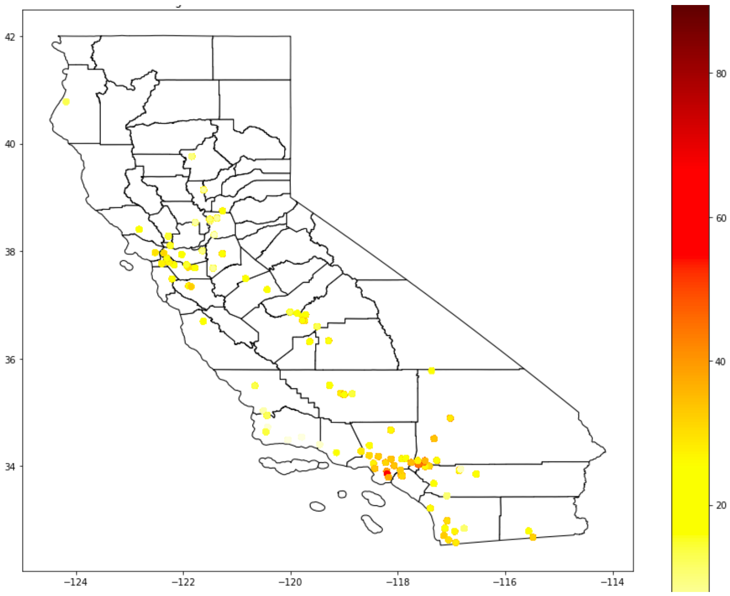


Figure 1: NO2 concentration level across California in 2020

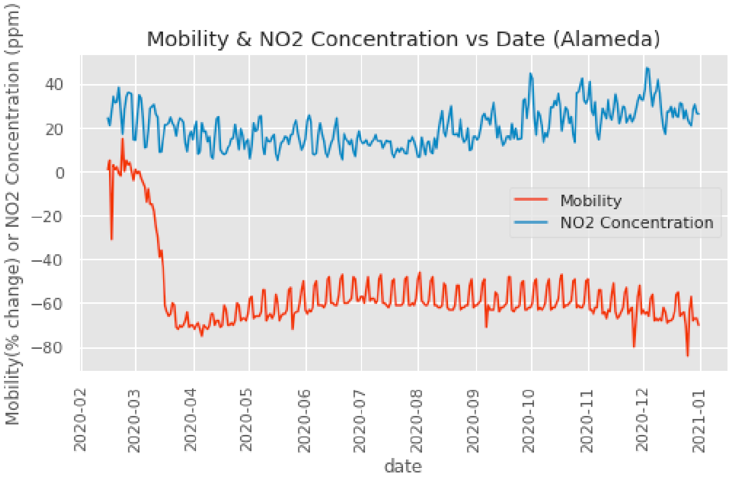


Figure 2: Mobility & NO2 Concentration vs Date in Alameda county

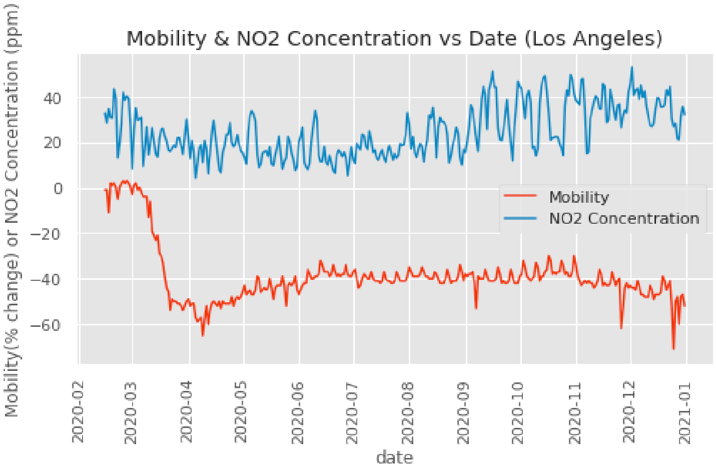


Figure 3: Mobility & NO2 Concentration vs Date in Los Angeles county

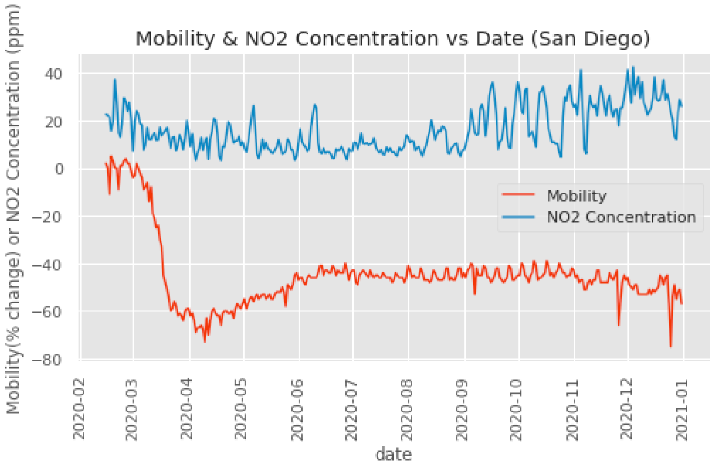


Figure 4: Mobility & NO2 Concentration vs Date in San Diego county

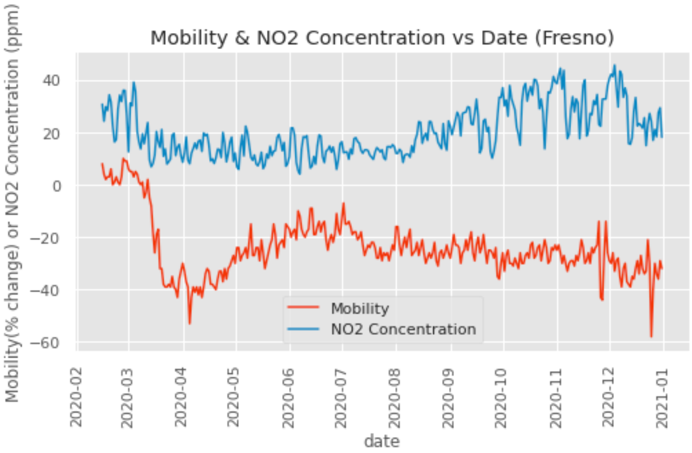


Figure 5: Mobility & NO2 Concentration vs Date in Fresno county

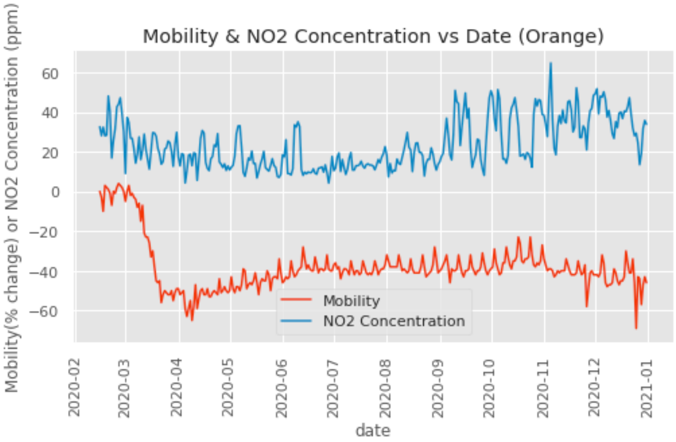


Figure 6: Mobility & NO2 Concentration vs Date in Orange county

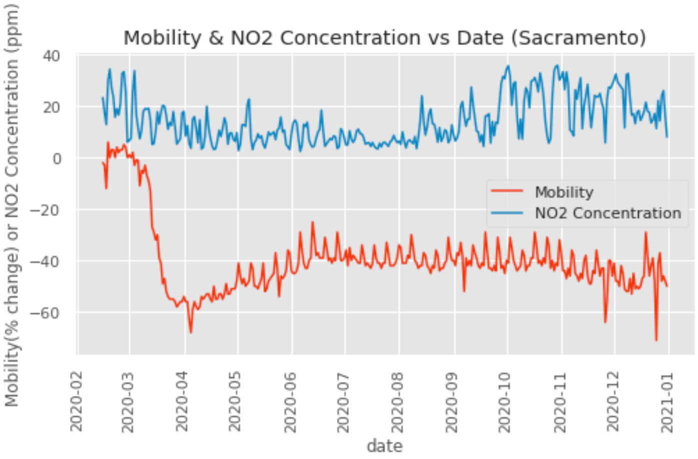


Figure 7: Mobility & NO2 Concentration vs Date in Sacramento county

For the other research question we proposed, we visualized covid prevalence and vaccination data for Alameda County in 2021. Covid cases showed a general decreasing trend in 2021 with the exception of August and December with new variants spreading globally. The distribution of doses of vaccination is bimodal with two peaks in May and December. We then combined mobility data with covid policy, and as shown in our boxplot, when stay at home policy was enforced, social mobility decreased.

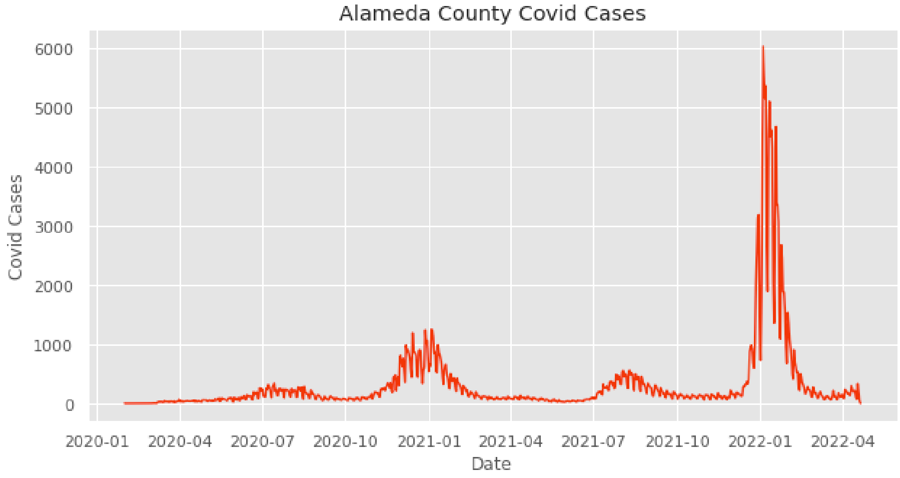


Figure 8: Alameda County Covid Cases

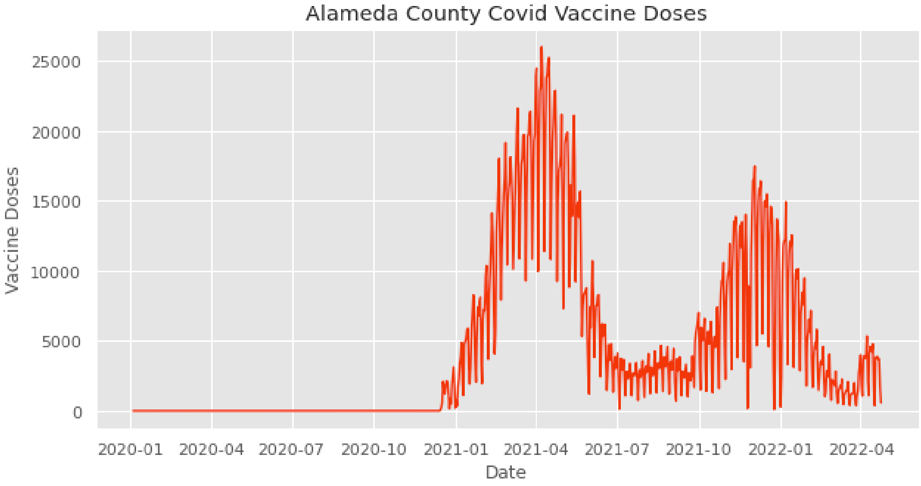


Figure 9: Alameda County Covid Vaccine Doses

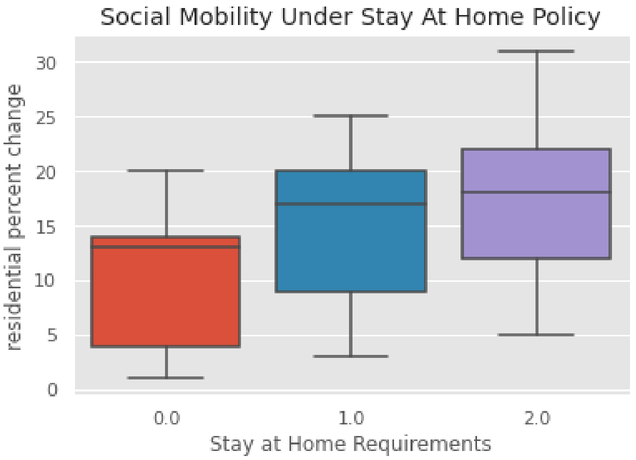


Figure 10: Social Mobility Under Stay At Home Policy

We selected only California data from the 2020 mobility dataset and kept only the county, date and transit percent change column because we wanted to focus on the human mobility on transit. Next, we created six different mobility datasets for six different counties which were Alameda, San Diego, Los Angeles, Fresno, Orange and Sacramento. For the NO2 dataset, we also created six different datasets for these counties and only kept the data, daily max 1-hour NO2 Concentration and county name column. After that, we used the date column as a key to merge the county mobility dataset and county NO2 dataset together for all these counties. Similarly, we filtered covid cases and vaccination data to Alameda County in 2021 and made our visualization. To investigate potential correlation between covid policy and social mobility, we merge the two dataset on dates and make visualizations from the combined dataset.

The red line in Figure 2~7 visualizes the annual trend and provides solid evidence of the pandemic's influence on mobility. After March 2020, there is a dramatic decrease in mobility, and such a result provides a basis to study whether there is synchronization between mobility and pollution. The blue line in Figure 2~7 shows the 2020 annual trend of Daily Max 1-hour NO2 concentration in six counties. From the NO2 line plot, we could also see that the NO2 concentration decreased around these months. This informs us to decide whether potential opportunities exist to examine the causal relationship between NO2 concentrations and transportation mobility. That graph also assists us in carrying out multiple hypothesis testing to analyze whether the regression coefficient is significant or not. (Null hypotheses: beta = 0 , Alternative hypothesis beta 0).

The last set of figures (Figure 8~10) is used to investigate potential causal inference between covid policy and social mobility. With covid policy being the treatment and mobility being the outcome, we visualize their relationship using a boxplot. However, since covid prevalence might be a confounder in the causal relationship, we also visualize covid cases and vaccination status in the specific region and the certain time period that we are focusing on.

4. Inference and Decisions

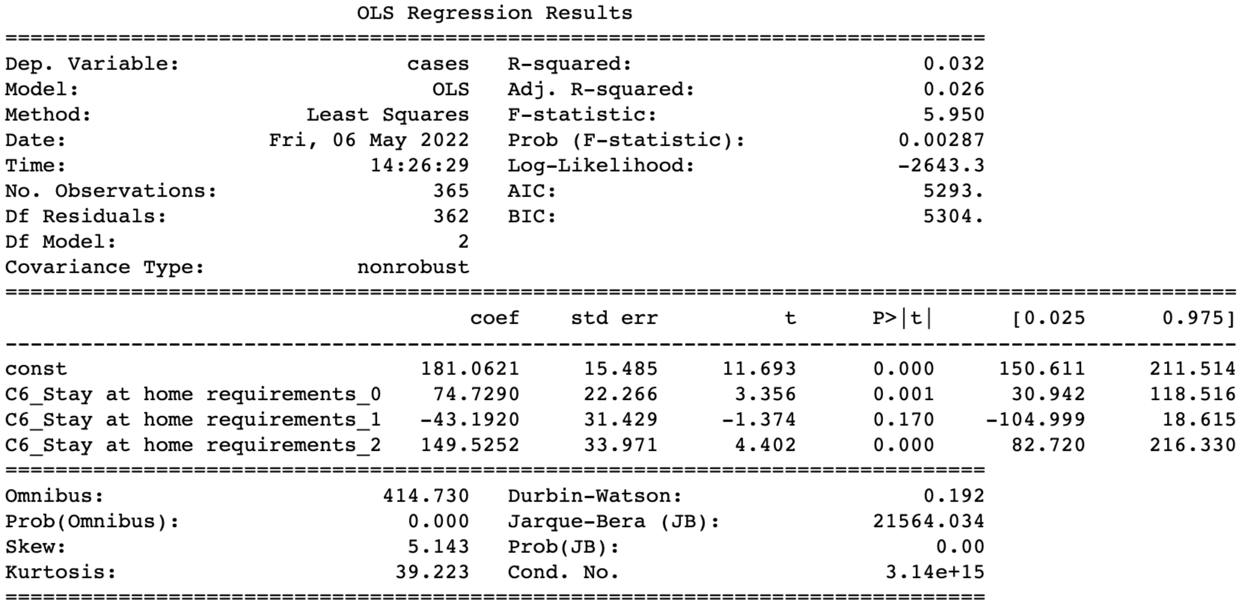
4.1. COVID-prevalence and mobility

In our implementation of causal inference, covid cases in Alameda County in 2021 is the treatment and workplace percent change in mobility is the outcome. Confounders in our model setup include vaccination status, unemployment rate, and other economic indicators such as GDP. The unconfoundedness assumption does not hold in our model, so we will have to adjust for confounders.

First, We use instrumental variables to adjust for confounders. Specifically, we choose the stay at home policy, which is a categorical variable measuring the intensity of the policy implementation, as the instrumental variable. The instrumental variables are able to isolate some of the confounding variables, such as vaccination, but are still correlated with some other confounders. There are no colliders in the dataset.

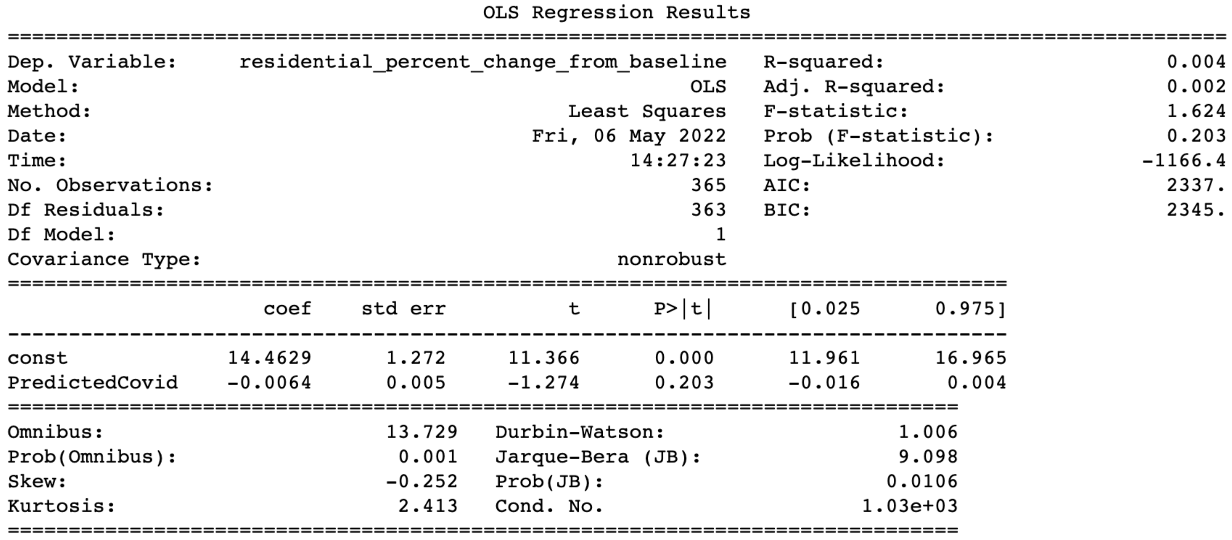
To implement our study, we use instrumental variables to perform two stage least squares. Stage 1 is to predict treatment variable Z_hat from instrumental variable W. Stage 2 is to predict target Y from predicted treatment variable X_hat. In Stage 1, since our instrumental variable is categorical, we first one-hot encode the stay at home requirement variable and then fit an OLS regression. The result of two stage least squares is shown in Table 1 and 2 below. The coefficient is -0.0064, which implies a lack of causality between our treatment and outcome.

Table 1: OLS Regression Results from Stage 1



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  cases   R-squared:                       0.032
Model:                            OLS   Adj. R-squared:                  0.026
Method:                 Least Squares   F-statistic:                     5.950
Date:                Fri, 06 May 2022   Prob (F-statistic):            0.00287
Time:                        14:26:29   Log-Likelihood:                -2643.3
No. Observations:                 365   AIC:                             5293.
Df Residuals:                     362   BIC:                             5304.
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                            181.0621     15.485     11.693      0.000     150.611     211.514
C6_Stay at home requirements_0    74.7290     22.266      3.356      0.001      30.942     118.516
C6_Stay at home requirements_1   -43.1920     31.429     -1.374      0.170    -104.999      18.615
C6_Stay at home requirements_2   149.5252     33.971      4.402      0.000      82.720     216.330
==============================================================================
Omnibus:                      414.730   Durbin-Watson:                   0.192
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            21564.034
Skew:                           5.143   Prob(JB):                         0.00
Kurtosis:                      39.223   Cond. No.                     3.14e+15
==============================================================================
```

Table 2: OLS Regression Results from Stage 2



```
                                 OLS Regression Results                                 
========================================================================================
Dep. Variable:     residential_percent_change_from_baseline   R-squared:                       0.004
Model:                                                  OLS   Adj. R-squared:                  0.002
Method:                                       Least Squares   F-statistic:                     1.624
Date:                                      Fri, 06 May 2022   Prob (F-statistic):              0.203
Time:                                              14:27:23   Log-Likelihood:                -1166.4
No. Observations:                                       365   AIC:                             2337.
Df Residuals:                                           363   BIC:                             2345.
Df Model:                                                 1                                         
Covariance Type:                                  nonrobust                                         
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           14.4629      1.272     11.366      0.000      11.961      16.965
PredictedCovid  -0.0064      0.005     -1.274      0.203      -0.016       0.004
==============================================================================
Omnibus:                       13.729   Durbin-Watson:                   1.006
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                9.098
Skew:                          -0.252   Prob(JB):                       0.0106
Kurtosis:                       2.413   Cond. No.                     1.03e+03
==============================================================================
```

One limitation is that, while the instrumental variable we chose is able to isolate some of the confounded variables, there are still other confounders such as economic performance under the pandemic, which significantly weaken the causality between our treatment and outcome. Another reason for the weak causal relationship is due to the fact that our instrumental variable is a categorical variable, which significantly increases the bias in our model. Having more data on measuring the intensity of the lockdown policy would be useful as using a categorical instrumental variable was a limiting factor for our model. Also, additional data on economic performance would be useful since indicators of the economy such as GDP and unemployment rate could provide us some insight into the covid prevalence.

4.2. Mobility and NO2 Concentration

Because our research question is to analyze whether human mobility in public transportation affected the NO2 concentration in six different counties. We plan to apply six logistic regression models and test whether each coefficient of the regression models is equal to zero or not. Since there will be six hypothesis testing that focused on each individual county, it is important to apply multiple hypothesis testing techniques to control the family-wise error rates and false discovery proportion rate.

From the above hypothesis testing, we want to test whether the coefficient of the logistic regression between our dependent variable (Average daily NO2 concentration) and the independent variable (transit stations percent change) is zero or not in six counties respectively. We would reject the null hypothesis if the test statistics suggest that the coefficient isn't zero. Such hypothesis testing makes sense because we are uncertain about whether correlation between two variables is negative or positive, and two-sided t-test would be preferred in this case.

We choose the Bonferroni Correction and the Benjamini-Hochberg correction to modify the p-values from our above five listed hypothesis testing. From the Bonferroni correction, we want to control the probability of any of the five tests that has a false positive (family-wise error rate). And we also want to control the proportion of discoveries that were wrong in our particular decisions by averaging over the randomness in the sequence of p-values for Benjamini-Hochberg correction.

When we go through the six hypothesis testing across the six interested counties, we find out that Los Angeles, Sacramento and San Diego county have significant effects that are influenced by the transit stations percent change, which suggest we should reject the null hypothesis of zero coefficients in these three counties. Whereas the three remaining counties have indicated a non-significant correlation, such results fail to reject our null hypothesis. The summary statistics of each logistic regression model is listed as below.

Table 3. Logit Regression model Summary Table.

	Coefficient	Standard Deviation	Z-score	p-value	Lower: 0.025	Upper: 0.975	Reject or not
Sacramento Mobility % change	0.0239	0.004	6.518	0.000	0.017	0.031	Rejected
Fresno Mobility % change	0.0045	0.005	0.894	0.371	-0.005	0.014	Not rejected
San Diego Mobility % change	0.0142	0.003	4.692	0.000	0.008	0.02	Rejected
Alameda Mobility % change	0.0002	0.002	0.072	0.943	-0.004	0.005	Not rejected
Los Angeles Mobility % change	-0.0098	0.003	-2.915	0.004	-0.016	-0.003	Rejected
Orange Mobility % change	-0.0042	0.003	-1.266	0.206	-0.011	0.002	Not rejected

Notes: 1) The dependent variable is the NO2 concentration. 2) The significance level is 0.05 for each hypothesis testing.

After applying both of the error rate controlled methods, we find out that either implementing Bonferroni or Benjamini-Hochberg correction for multiple hypothesis testing results align well with our previous testing outcome as indicated in Table 3. Bonferroni correction controls for family-wise error rate for all six tests and B-H are explicitly designed to control error rate of the three individual discoveries. The following graph could visually show our multiple hypothesis testing results.

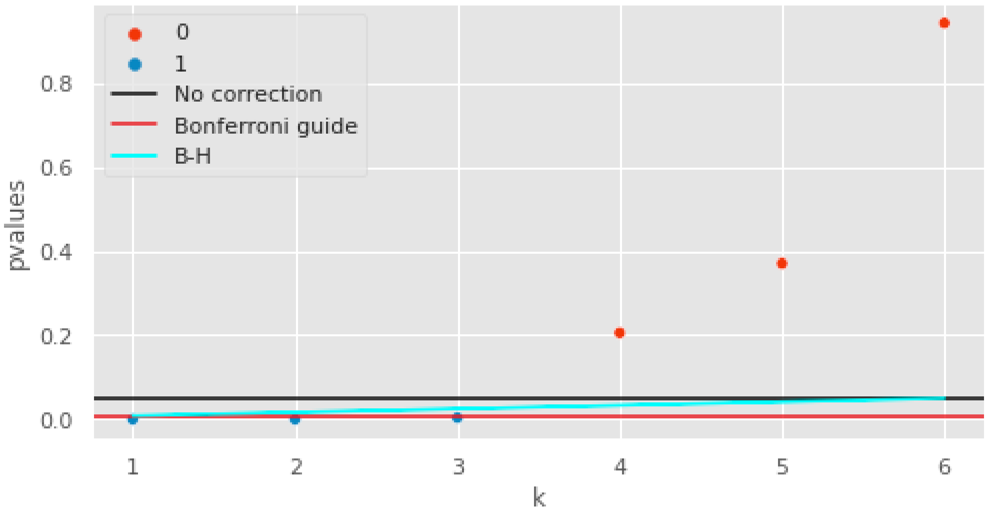


Figure 11: P-values vs k (number of test)

There are three discoveries that remain significant after we apply two correction procedures, which means that Los Angeles, Sacramento and San Diego county have significant effects that are influenced by the transit stations percent change.

Table 4. Model Performance for Each County.

Model Performance	Accuracy	Recall	F1-score	p-value	Reject or not
Sacramento	0.7113	0.1379	0.2222	0.000	Rejected
Fresno	0.4948	0.1296	0.2222	0.371	Not rejected
San Diego	0.6598	0.1622	0.2667	0.000	Rejected
Alameda	0.4329	0.28	0.3373	0.943	Not rejected
Los Angeles	0.7010	0.9855	0.8242	0.004	Rejected
Orange	0.5773	0.8571	0.7007	0.206	Not rejected

Notes: 1) The significance level is 0.05 for each hypothesis testing. 2) All of the above data could be referred to the confusion matrix in the jupyter notebook.

The table above shows the logistic regression performance of each county. We also include a series ROC and precision recall curve in our jupyter notebook to visualize our result. We notice that some counties have a really low recall score. This means our model has a high number of false negatives for that county. This may be because of the imbalanced class we have. Some counties also have a low accuracy score. This may be because we miss out some important confounders in the model or the correlation between mobility and NO2 is not obvious in that county. We also found that our model predicted the result on Los Angeles county pretty well. However, we predict none of the true negatives and only one false negative. This may be because the dataset that we use to train is too small. Overall, larger counties seem to have a better performance and successfully reject the null hypothesis. This may because larger populations will have a more significant change in mobility and the impact on NO2 concentration is more obvious in these counties.

Because we use two-sided t-test to validify our hypothesis testing, when we encounter negative coefficients from the logistic regression model, we could still reject the null hypothesis even though it is hard to explain from the perspective of environmental science domain knowledge. Also, there might be potential confounders at the county level that influences the correlation between our target variable and the independent variable. Thus, we might need to further consider exploring and including these confounding variables into every hypothesis testing procedure. This will also help to avoid the problem of p-hacking as each of the models are consistent with the same sets of controlling variables.

Since the monitor data points are unevenly distributed across different counties, there might be an imbalanced amount of input data into every logistic regression model based on a specific county. Therefore, if more data could be provided in each county that will enhance the accuracy of the logistic regression. And we could carry out more hypothesis testing with more counties, since currently some of the counties do not have enough data to allow us to implement a corresponding logistic regression model.

5. Conclusion

For the first question, we investigate if there is a causal relationship between covid prevalence and mobility. We use covid policy as the instrumental variable and conducted two stage least squares. The resulting coefficient is -0.0064, indicating a lack of causality between covid prevalence and social mobility. R squared of our OLS regression is 0.004, also implying an unsatisfactory goodness of fit. For the second question, We analyzed how public transportation affected the NO2 concentration in Alameda, San Diego, Orange County, Fresno, Sacramento, and Los Angeles. Using the technique of logistic regression and multiple hypothesis testing, we successfully identify two significant positive relationships between mobility and NO2 concentration in San Diego and Sacramento, and we also discover a significant negative correlation in Los Angeles. All of the remaining tests have failed to reject the null hypothesis. We also conclude that all of the three discoveries remain valid after we correct the family-wise error rate and false discovery rate.

With regards to the results of the second question, in terms of spatial generalizability, since the result focuses with a speciality on six counties of California, we could consider applying the same mechanism to other counties if enough data could be provided. However, it might be difficult to apply to other states as we explain before that spatial confounders could be influential in challenging the robustness of the hypothesis testing results. In terms of temporal generalizability, our data focus on 2020 that experience in the middle of the pandemic, which provides us a special opportunity to capture a rapid decrease in the mobility and thus to correlate these points with the local NO2 concentration level. However, if we experience an economic recovery in the post-pandemic era that is likely to cause a sharp increase in mobility, this would be harder to explain by our current model.

With a serious destruction of the pandemic, policy makers in developed countries have a unique opportunity to integrate further land-use planning and transport to encourage low-emission, low-motorized mobility. We suggest that policies should focus on compact, connected, and efficient forms of urban development to encourage people to shift from traditional vehicles that lead to high levels of emission into walking, cycling and other modes of clean vehicles to reduce the NO2 pollution.This specifically applies to places with bigger geographical range (i.e. San Diego, Los Angeles, and Sacramento) as we discovered in our second research question. Moreover, since the pandemic is still ongoing, we suggest public health officials to closely monitor covid prevalence and social mobility, sending covid risk warnings correspondingly to stop the spread of the pandemic.

Our data, being from multiple datasets, need to be merged. Initially, we merged the mobility dataset and the NO2 concentration dataset by taking the date as a key. By merging these two datasets, we can create a single plot for both these two variables. This makes it easier to visualize the changes between mobility and NO2 concentration with interaction to dates. First, the Google Community Mobility Report is being reported in the middle of a pandemic to allow the public health scientist for better policy making. However, it fails to trace back the data before 2020 as it sets a baseline on a certain date and captures the fluctuation from that date. And the mobility data could be hard to be correlated with the historical air pollution data. Second, as we explained before, the monitor environmental pollution data is unevenly distributed. This might cause difficulty when we apply the same methods to other places.

Based on our result, mobility does have association with NO2 concentration in some of the counties. In the future study, if people want to analyze the change of NO2 concentration with some other factors, transit mobility could be one of the confounding variables they need to consider. This project could be reproducible for other counties or states, and this could also help policymakers decide which counties should apply certain regulations on reducing the usage of transportation. For the first research question, even though our analysis does not produce the desired results, future study could still build on our result when more research on relevant fields is available.