Data 102 Final Project

Suhang Xiang, Ash Feng, Daniel Bostwick

I. Data Overview, Questions and EDA

A. *Data Overview*

Our project required a few different datasets collected across different sources. Our main datasets were census data collected across the whole nation so there were no exclusions. We did generate three small datasets for California, Illinois, and Kansas. Our additional data was to provide insight on political shifts in relation to changes in mobility. We don't know how specific participants were affected by this data but we do know that the data we had was quite granular in that we had access to specific counties and their respective vote counts. For the transportation data, the data was separated by month and year but was an aggregate for the whole country. Our data was rather intuitive and there weren't any forms of bias or errors nor was there any differential privacy.

We wish we had more confounder variables but the data lacked a sufficient amount. We wish we had access to more senate data at the county level so we could have better regression models. For both questions there were a lot of data missing so we removed those columns and rows to avoid NaN errors. We had to do a lot of data cleaning to make our data not only legible but also understandable and useful for causal inference and regression analysis. Please reference to the file data102_project.ipynb in submitted zip file via Gradescope.

B. *Research Questions*

All matters dealing with this section are specifically outlined in the their respective area.

Question 1: Did the change in mobility during COVID (2021) affect voting patterns in 2022 election?

We will use a Neural Network and two different sigmoid GLMs because these methods provide good binary outputs for our predictions. The limitations of these methods is finding an ideal prior and likelihood distribution which if bad will give us a bad result in our posterior.

Question 2: During the pandemic, will the changes in the flow of people flying in Domestic(U.S.) affect the country's investment in the aviation industry?

We want to study the causal relationship and find the best explanation of inference for the question above, so causal inference method is good choice. Some limitations are possibly not having enough confounding variables.

C. *EDA*

1) : In Figure 1, it is very interesting that the mean of the Republican data is very close if not sitting at zero. However, the mean of the Democratic data is sitting just left of zero. By left, this is indicating that there is negative mobility from the baseline in the Democratic side. We chose to show only two of the 32 states (66% of all states due to senate election cycles). In our observations among the 12 states we actually visualized, the distributions were about the same. So our selection of Texas and New Jersey were chosen arbitrarily to give a visualization of the trends we saw of the overall data. On the right hand side, we included the total 32 states and we saw that the overall trend among all states is fairly similar to that of the individual state. The workplace variable is our only outlier variable where the data isn't exactly like the other two variables. Although the data isn't exactly similar with other two, it is similar in the overall trend of the each state. There are more Republican counties in the US than there are Democratic and this is shown US plot. If people from red states didn't have to change much about their life in terms of mobility, then they might not be willing to vote for someone outside of their party. We see high volumes of shifting from blue states which might entice them to change parties due to highly restrictive policies. Whereas people from red states might change party affiliation due to the lack of restrictive policies which led to higher contraction rate of the virus leading to higher death per capita.

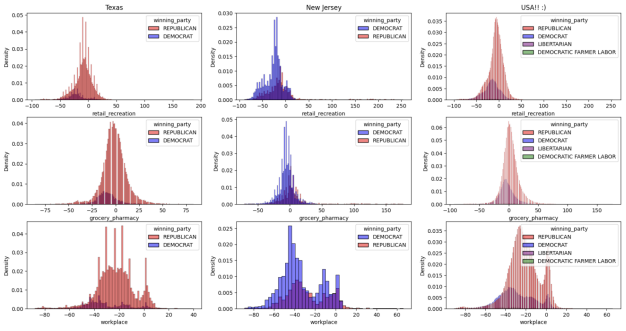

Fig. 1. Histogram of numerical variables with a hue on party affiliation.

2) : In Fig.2 we see during the COVID pandemic, local government investment in air has actually increased significantly compared to normal years.

3) : In Fig.3 as we common sense know, the COVID epidemic will inevitably lead to a reduction in domestic traffic, from the plot we can intuitively understand how much domestic traffic have dropped compared to usual.

4) : In Fig. 4 we see there is indeed a somewhat linear relationship between the two variables. However, we may not conclude that there is a causal relationship between the two. In order to determine whether there is a causal relationship between the two, we will use the causal inference method to analyze in the next stage of the project.

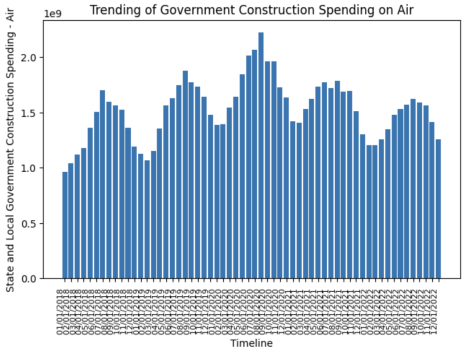

Fig. 2. Plot of Trending of Government Construction Spending on Air(2018-2022).

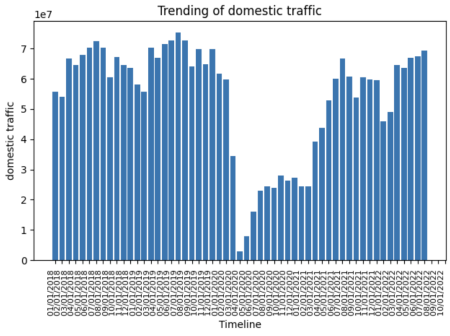

Fig. 3. Plot of Trending of domestic traffic (2018-2022).

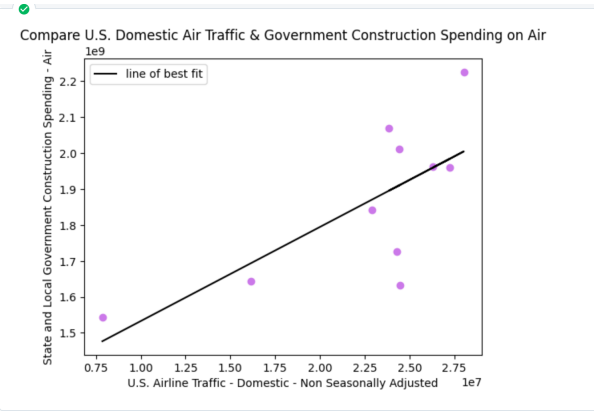

Fig. 4. Fit line of domestic traffic and air spending of lag 2 months.

5) : In Fig. 5 In order to intuitively understand the state and local government air investment in different periods, we created a categorical variable based on the data to compare the investment in different periods. This data comparison will be of great help to our subsequent data analysis work.

II. Causal Inference

This part will analysis Question 2: During the pandemic, will the changes in the flight traffic in Domestic(U.S.) affect the state and local's investment in the aviation industry in U.S.?

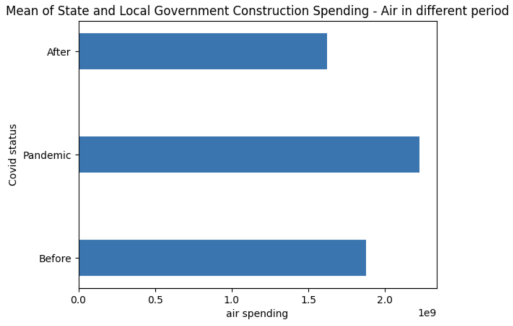

Fig. 5. Mean of State and Local Government Construction Spending - Air in different period.

A. *Methods*

*1) **Treatment and outcome:**:* In order to see whether domestic traffic of air would cause local government investment on air, we use domestic traffic(U.S.) in air as treatment, and use state and local Construction Spending in air as outcome. Since domestic traffic of air is a continuous variable, in order to simplify the analysis of Causal Inference, we transform this continuous variable into a dummy variable by an easy method. Since domestic traffic has a very significant difference during the pandemic, we transform domestic traffic into a dummy variable. We set the treatment "1" for the time interval when domestic traffic decreases significantly and the control "0" for the normal traffic during the normal time period.The time period for this treatment is March 2020 - March 2021 as seen in the EDA analysis above.(March 2020 Began to decline significantly and reached lowest peak in April 2020).

*2) **Confounders and unconfoundedness:**:* We will use international traffic as confounder. First, we will choose the international traffic variable as one of our confounder. It is obvious that international traffic will simultaneously affect both the treatment and the outcome variable. This is because a large number of passengers arriving in the United States from all over the world on international flights, whether for business or tourism, will travel through multiple cities within the United States and most will choose air travel as their mode of transportation. This inevitably affects domestic traffic which is the treatment. In addition, the place-based aviation investment, which is the outcome variable, is closely related to the magnitude of traffic, whether domestic or international. This is why we chose international traffic as the confounder.

*3) **Adjust for confounders:** :* To adjust for confounders, we use the outcome regression method to see the result. Also, we use Inverse Propensity Weight(IPW) method to compare the effectiveness of unconfoundedness with the outcome regression.

*4) **Colliders:** :* If a variable is caused by the treatment variable and outcome variable, we call the variable as collider. In order to determine whether there are variables that are affected by domestic traffic of air and at the same time affected by local government investment on air, we checked the dataset, but there was no obvious evidence of such variables.

*5) **DAG draw:** :* See the Fig11.

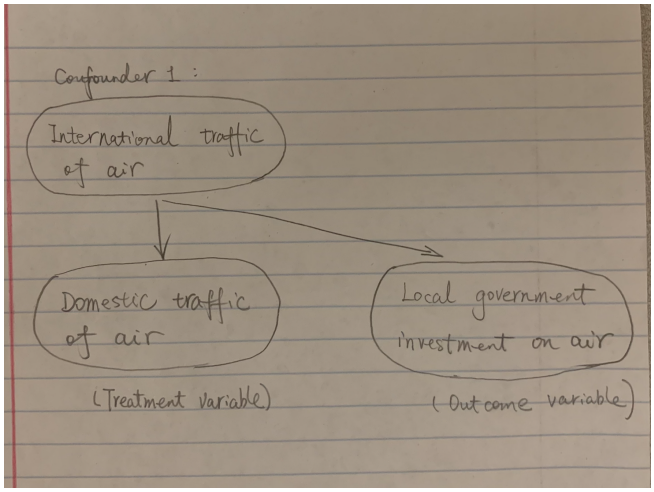

Fig. 6. DAG draw for the causal Inference.

B. Results

*1) **Summarize:** :* We use the "statsmodels" library to do the Outcome Regression. The result shows that the ATE is 3.43e+08 dollars. However, when we do Bootstrap,our 0.95 confidence interval ranges from 1216636972.0 to 1686138487.0, and the observation is out of the range. Assuming that the confounder we selected is correct, the results show a lack of clear causal relationship between the domestic traffic and the local investment on air. So, even there is some linear relationship between the domestic traffic and the local investment on air, we cannot claim that domestic traffic cause the change of local investment on air.

OLS Regression Results

Dep. Variable:	air spending_lag 2 mons	R-squared (uncentered):	0.980
Model:	OLS	Adj. R-squared (uncentered):	0.979
Method:	Least Squares	F-statistic:	738.2
Date:	Wed, 03 May 2023	Prob (F-statistic):	2.90e-38
Time:	23:32:55	Log-Likelihood:	-991.07
No. Observations:	48	AIC:	1988.
Df Residuals:	45	BIC:	1994.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
treatment	3.43e+08	1.23e+08	2.797	0.008	9.6e+07	5.9e+08
inter traffic	-24.4525	18.393	-1.329	0.190	-61.497	12.592
hvm	0.0062	0.001	10.601	0.000	0.005	0.007

Omnibus:	0.620	Durbin-Watson:	0.607
Prob(Omnibus):	0.733	Jarque-Bera (JB):	0.589
Skew:	0.251	Prob(JB):	0.745
Kurtosis:	2.792	Cond. No.	9.70e+11

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 9.7e+11. This might indicate that there are strong multicollinearity or other numerical problems.

Fig. 7. Outcome Regression result

*2) **Uncertainty:**:* The causal inference is based on the time period of the covid pandemic. Because the time of the pandemic is between March 2020 and May 2022, the amount of data in this interval is very small. This adds to the uncertainty in the data due to the small sample size.

C. Discussion

*1) **Limitations of methods.:** :* Due to the sharp decline in domestic traffic, which occurred during the pandemic period from March 2020 to May 2021, the sample size based on monthly data is small, making the precision of the Bootstrap simulation relatively rough. This is the first limitation. Additionally, we found that the selection of confounders has a crucial impact on the experimental results. Once different confounders are chosen, the results of the experiment will change significantly. The selection of confounders requires rich professional knowledge and rigorous logical reasoning on this issue. In reality, inferring local government investment from domestic air traffic is a complex causal inference. Since the original dataset has few confounders related to our study of causal issues, we need to search for potential confounders from additional datasets. Our research involves complex macro factors (such as government budgets) which may hidden confounders. Thus, the limitations of our current expertise make it difficult to discover these potentially crucial confounders. This is another limitation of our causal study. Despite these objective limitations, we believe that exploring the causal relationship of this issue is meaningful because there is indeed some linear correlation between flight traffic and government investment, as explained in the previous EDA discussion. What we need to do is to determine whether this correlation has a strong causal relationship.

*2) **Additional data**:* We introduce additional dataset of covid vaccination to support the study of this one causal inference question. Since we are looking at the causal relationship between domestic air traffic and local investment on air during the pandemic, which is based on the pandemic timeline. Therefore, we introduce this additional dataset to corroborate and define the pandemic timeline. But, this additional dataset we only use for the EDA study. During our research, we expect to find more confounder factors and hope to find more additional dataset based on those confounder variables to support our study. However, as in the analysis of confounders above, it is difficult to find more corresponding confounders for this research topic we set although we expect to cite more additional datasets.

*3) **Confident of the causal relationship**:* We don't have confident with the causal relationship between our chosen treatment and outcome, which is domestic air traffic decreasing in pandemic cause more local investment on air lagging 2 months although these is a obvious linear relationship. One reason is that the observed value of ATE is out of the confident interval which we estimated with Bootstrap. Another reason is that we lack to find more significant confounders because the background of this research topic is so macroscopic that the real confounders hidden behind need more professional expertise domain to reveal.

III. GLM AND NON-PARAMETRIC MODEL

	state object	county object	retail_recreation fl	grocery_pharm...	workplace float64	winning_party o...
0	ALABAMA	AUTAUGA	5.0	7.0	-4.0	REPUBLICAN
1	ALABAMA	AUTAUGA	0.0	1.0	-4.0	REPUBLICAN
2	ALABAMA	AUTAUGA	8.0	0.0	-27.0	REPUBLICAN
3	ALABAMA	AUTAUGA	-2.0	0.0	2.0	REPUBLICAN
4	ALABAMA	AUTAUGA	-2.0	0.0	2.0	REPUBLICAN
5	ALABAMA	AUTAUGA	-8.0	-3.0	1.0	REPUBLICAN

Fig. 8. Table used for predicting party affiliation.

A. *Methods*

*1) **Overview**:* Our initial attempt at answering this question was to use the columns retail_recreation, grocery_pharmacy, and workplace to regress on winning_party. We chose these columns to use for our models because as seen in Fig. 2 above, each of these parameters follows a relatively normal distribution. Also, each of the quasi-Normal distributions have a separate quasi-Normal distribution according to the known party affiliation at that time. The state level histograms show a fairly noisy distribution but the overall US histograms are pretty smooth over their respective surfaces. This means that there would be potentially separated hyperplanes distributed throughout the data hopefully making for an easier or more accurate prediction on party affiliation.

This first attempt was at the county level and we wanted to guess as far down as county majority of the senatorial party affiliation to see if, based off of mobility or lack thereof, we could guess which party each county would vote for. Our logic was that even though there were some federal mandates and many state mandates that were a little different according to which state you resided in during the COVID-19 pandemic, there were also many county mandates within each state. These mandates varied in strictness and we think could have affected the way a specific county would have voted in the 2022 election.

Later in this paper we explain some of the troubles we encountered with this specific approach, but our final method of predicting party affiliation was very similar. We decided to regress over county affiliation, but with limited data at our disposal we ended up only testing on three states: California, Kansas, and Illinois. We chose these states on purpose because these states represent the fundamental differences seen in voting populations around the U.S. California is a sold blue state, Illinois is a purple state, and Kansas is a solid red state. The data from our training set and these three testing sets were the same so we did not need to worry about changing anything to the existing format of the training dataset but we did need to binarize the counties' party affiliation to make sure we trained our models on the numerical data representation of party affiliation. To test these three states we used a Neural Network, Logistic Regression Model, and a Sigmoid inverse link function with assumed Uniform distributions on the data.

*2) **Neural Network**:* Our first question in picking out a non-parametric model was to decide which type of model could most easily separate data and predict the correct binary output. We immediately decided that a Neural Network might be pretty good for a non-parametric model as we have some experience with Neural Networks and we know that they are a powerful tool for predicting binary outputs. We came up with a classic Neural Network model solving for our optimal values by implementing SciKit Learn's MLPClassifier() module. Our hyperparameters include a solver Adam that extends stochastic gradient descent which is really helpful with these types of networks and is implemented over five epochs. Our network consists of four layers, the two hidden layers are of size 100 and 50, and the output layer is binary outputs. To test the accuracy of our outputs we utilized the score() function from SciKit Learn and compared it to every state. We used the accuracy from the Neural Network as the baseline prediction accuracy for our other two models.

*3) **Frequentist GLM**:* We first came up with our Frequentist model by examining the desired output and because we aimed at achieving good binary predictions, predicting county level party affiliation, we knew that a Logistic Regression model would be a great model to fit to our data. As mentioned before we trained our model on three columns of data: the mobility change from a set baseline in retail and recreation, grocery and pharmacy, and workplace and set our training and validation sets up into an 80/20 split. Both our Frequentist and Bayesian models make use of a Sigmoid function but for our Frequentist model we don't know the values of β that would give our model the most optimal coefficients to train on. We made use of the SciKit Learn LogisticRegression() module to train our data on. We made no assumptions on the data for this except for the fact we know that our output is binary so it is Bernoulli, but that fact isn't applicable in this instance. For evaluation we scored the training set's predictions against that of the true values in the data set. Then, as mentioned above we tested our model on the three different states California, Illinois, and Kansas.

*4) **Bayesian GLM**:* Bernoulli plays a large factor in our Bayesian GLM. Moreover, the Likelihood function we use for our Bayesian model is the Bernoulli Likelihood. For this regression model, we decided on using PyMC3 library to formulate our values of β that will be used in the final part of this process. Our final implementation is designed as follows: Our dataset was the same as before but even though we saw that our columns of data from 2020 followed a quasi-Normal distribution, we actually applied a highly naive Uniform distribution to the data for training purposes. As seen in Figure 2 above, the data is spread out roughly between (-100, 100) along the x-axis. We used this observation to gather a naive estimation of the what the parameters would be for a Uniform distribution. Our decision for this was due the the following two reasons, we saw that our naive Normal distribution was failing horribly and when we took a moment to step back from our original observations, we realized that every county has an equally likely chance of being either

Democrat or Republican and that our above graph shows the density of those observed outcomes.

With that figured out we then used a Normal distribution with a mean of 0 and a standard deviation of 1 for our intercept term. Our logic for the intercept's distribution is that there is no known distribution for this term so by statistical convention we should assume normality. As mentioned before, the Likelihood for this model is Bernoulli but we need to setup our Likelihood function properly. To set up our model properly we used the following equation in our model:

$$\beta^T x_i = \beta_{int} + \beta_{rr} \cdot rr_i + \beta_{gp} \cdot gp_i + \beta_{wp} \cdot wp_i \quad (1)$$

where int is integer, rr is retail and recreation, gp is grocery pharmacy, and wp is workplace. After setting this up we finally can run the PyMC3 module and where we use our observed y values from the 2020 dataset to predict our β values on. We receive our estimated values for β and since we know we are using a Bernoulli Likelihood we can plug in our new values for β that came from the assumed prior and Likelihood to the following equation:

$$\sigma(\beta^T x_i) = \frac{1}{1 + e^{-\beta^T x_i}} \quad (2)$$

Now that we have the β values of our posterior distribution, we can plug in those values to our 2022 states and find a prediction to help us answer our original question.

B. *Results*

*1) **Neural Network**:* The states results were as follows:

California: 54.46%
Illinois: 74.77%
Kansas: 91.9%

These results surprised us at first until we analyzed the states more closely. Each state has a specific quality among them that make them quite different when running these types of regressions models. California is a solid blue state but the counties in California are surprisingly quite equal in nature. There are many densely populated red and blue counties so they hold similar weights in overall elections. Apparently what keeps California so blue is the Los Angeles and San Francisco areas, but other than that California is a rather purplish state. Illinois is an easier state to make observed distinguishes in but still there are some heavily populated blue areas peppered in between which make the state itself blue but the county vote a little easier to distinguish. Kansas is solid red for almost every county except for the county Kansas City resides in.

*2) **Frequentist GLM**:* The states results were as follows:

California: 47.98%
Illinois: 74.39%
Kansas: 91.68%

After seeing the results from the Neural Network, these results did not surprise us as much as we had already learned more about why these numbers might be true. There was little uncertainty with these results and as we chose to use a very direct approach to our Frequentist GLM, we knew that our model was very efficient. Later in the Discussion section we talk about the different hyperparameters we tried and how those effected the model.

*3) **Bayesian GLM**:* The states results were as follows:

California: 62.12%
Illinois: 66.24%
Kansas: 53.31%

This GLM was very unusual to us and we still aren't very confident as to why these results came out to be this bad. Although we see a clear increase in the prediction score of California's county voting pattern, the others decrease so significantly that we cannot presume that California is right either. What we found interesting in these results is that we used a very similar approach to our Frequentist GLM. Both the Frequentist and the Bayesian are implementing a Sigmoid function to evaluate their binary predictions. One uncertainty we had about our model is that our prior distributions were so random to us. Even though we explain above the logic that went into choosing a Uniform distribution within the range $(-100, 100)$, our first assumption about a Normal prior still made some logical sense to us. However, when we did implement a Normal prior on the variables x_i from our dataset, the prediction accuracy was below 20% for all of the states. We knew this was not accurate at all so that's when we definitively were set on a Uniform prior on all the x_i's in our dataset.

C. *Discussion*

*1) **Neural Network**:* Our mean loss of the five epochs was 0.485 with the min being 0.479 and the max being 0.497. The overall performance of the model predictions when compared to the testing set was 78.23%. This model ended up performing the best out of all of the other models. It's relative overall performance was not bad either, with only three data points we were able to muster an almost 80% accuracy rate on the validation set. Maybe if we used vote share percentages in our model our model might have performed better, but there is also a good chance of over fitting as well. The overall performance of this model was decent but not good, this model would probably be good to use if you wanted to get a feel for what kind of state you are looking at instead of trying to actually guess a county from a particular state. The results we rendered were definitely more of an analysis of the state's county population more than predicting the right county. This is seen in both the Neural Network and the Frequentist GLM.

*2) **Frequentist GLM**:* The coefficients for 2020 dateset were $[-0.01226893, -0.0107064, -0.01822611]$, and the calculated intercept was -1.81501736, and the validation set accuracy was 77.18%. This score is not too bad given the only three data points provided. This model's accuracy was slightly lower than that of the Neural Network in every regard but only by a small margin. Just as the Neural Network, the Logistic Regression model output some very similar accuracy scores on states. We originally tried using the vote share as well in this model but it overfit and saturated the data just in some instances and was completely drowned out in others. We did not use any hyperparameters within the LogisticRegression() module because they did not affect the outcome of the model.

*3) **Bayesian GLM**:* In Fig. 9, the coefficients for the 2020 dataset were $[-20.532, -25.398, -1.846, -0.224]$). These are the mean of the 1000 β values we generated. This is the worst performing model with scores well below the others. The major limitations we faced with this model is distribution of our data. It was very difficult to determine a good prior for our model to train over. We knew that our Likelihood was Bernoulli and that the output was going to be a Sigmoid inverse link function. We are still unsure as to why the model performed so poorly but it's a possibility that the variance was so large in each of the values of β that the mean was not a good representation of the overall data. Maybe if we used the median value of β among each β_i then we might have had a better result.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
intercept	-0.224	0.731	-0.812	0.988	0.364	0.279	4.0	4.0	4.26
rr_theta	-20.532	28.707	-43.286	28.777	14.296	10.946	4.0	11.0	4.03
gp_theta	-25.398	16.727	-40.101	2.694	8.330	6.378	4.0	4.0	5.00
wp_theta	-1.846	38.362	-41.593	46.293	19.104	14.628	4.0	11.0	4.03

Fig. 9. Bayesian GLM coefficients.

IV. Conclusions

*1) **Key findings** :* For the causal inference question, we don't find strong evidence that domestic air traffic decreasing in pandemic cause more local investment on air lagging 2 months.

For our predictions on county party affiliation, we found that our models are not very accurate in predicting. We probably need more data points for our model, because we only used those three quasi-Normal data points we were limited on the training our model could actually do. Since we eventually assumed a Uniform prior for our Bayesian model, we could have possibly implemented more of the columns of our original dataset.

*2) **Generalization of the result** :* For the causal inference question, after we did the research, we find the question we design is little bit narrow. more than study the relationship between the domestic air traffic and the local investment on air, it's better to study the relationship between domestic air traffic and domestic economic change because it is more generalization and significance in economic field.

For the GLM and non-paramteric question, we notice a very low accuracy rate overall but according to [3] there is only about a 60% accuracy rate among some of the best models when it comes to predicting results of an election.

*3) **Action** :* We recommend that the websites we got our data off of to provide access to tables (.csv, .xls, etc.) so we don't have to manually input them by hand which took a lot of time. This is a clear representation of the struggle data scientists have all around the world. When finding the major datasets we did use, MIT provided a full .csv file but it wasn't uploaded until a year and half after the election took place.

*4) **Data sources** :* For the causal inference portion we did use two datasets but their analyses wer done separately so there was no need to actually merge these datasets. However, for the GLM and non-parametric model we did have to merge quite a few datasets. We collected Senate data from 2020 and 2022 as well as Senate and state data from California, Illinois, and Kansas. These were all merged with our original 2020 and 2022 datasets about change in mobility during the COVID-19 pandemic.

*5) **Limitations** :* In the causal inference question study,there are two main limitations. First, it is difficult to find more confounders due to experts domain knowledge. it makes our confounders assuming lack and may not account deeper in our analysis. With more time to explore and better data mining skills, maybe we can find more meaningful additional dataset to support us to find more confounders. But for our project team of only three people, it is obviously unrealistic. Another limitation is that, the data (rows) is not big enough. This prevents the model(Bootstrap) from fully performing the simulation.

For the GLM and non-parametric model, the biggest limitation was access to good testing data. We were only able to really test our models on three states that we had to input data in by hand for. For every county in California, Illinois, and Kansas the county data was researched and annotated by hand to be able to conduct our regression models on on seen data.

*6) **Future studies** :* Although we did not find a clear causal relationship in the first question. However, it is meaningful to explore the relationship between flight traffic and local investment in air. Because there must be some kind of logical connection between the two. Changes in air traffic will inevitably have an impact on economic activity. At the same time, it will affect the local government's investment in air. Although this effect is not a direct causal relationship, it will help us discover the confouder factors behind it. Once we can determine the relevant confounder variable and have some market or decision-making intervention in it, it will help better decision-making for local investment. More importantly, through our research, we can broaden the subject of scientific research in the future. Perhaps it would be more meaningful to study the impact of changes in traffic variables on the overall economy than just looking at local investments in air. And the observation of traffic variables is not limited to air, but more extended to train and highway transportation.

*7) **Learn from the project** :* Through this project, we found that the ability to explore the raw dataset is very important. First of all, we must be able to understand the dataset, and discover valuable information from the known data, and understand what the existing data tells us. Only on this basis, we can design valuable and meaningful scientific research topics. After determining the scientific research topic, through the methods we learned in data102 (eg: GLM, Causal Inference, etc.), use logic and domain knowledge to further think about what additional dataset we need to support our arguments. In order to explore more additional supporting data as the basis of the research. Modeling and conclusions are then completed on the basis of comprehensive data.

V. Group Member Evaluation

1) ***Contributions :*** Ash Feng and Daniel Bostwick worked for the question of GLMs & nonparametric methods. Suhang Xiang worked for the question of Causal inference methods .

2) ***A report :*** We contacted another randomly assigned team member (Emir Karabeg) several times at the beginning of the project, but this person never contacted us and participated in the project throughout. So we actually only have 3 people working on this project. We hope that the data102 staffs can take into account the special situation of our group, and expect to give extra points in the results as appropriate. Thanks.

VI. Bibliography

References

[1] U.S. Senate Precinct-Level Returns 2020, https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc.

[2] Google: Daily Community Mobility Data. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ER9XTV.

[3] Laura Counts, 'Election polls are 95% confident but only 60% accurate, Berkeley Haas study finds', October 26, 2020.[Online], Available: https://newsroom.haas.berkeley.edu/research/election-polls-are-95-confident-but-only-60-accurate-berkeley-haas-study-finds/. [Accessed May 1, 2023]

[4] General Election - Statement of the Vote, Novermber 8, 2022 :: California Secretary of State. https://www.sos.ca.gov/elections/prior-elections/statewide-election-results/general-election-nov-8-2022/statement-vote

[5] Kansas Secretary of State — Election Results. https://sos.ks.gov/elections/elections-results.html

[6] Vote Total Search Election Results https://www.elections.il.gov/electionoperations/votetotalsearch.aspx