A7: Project Report

By Venkata Sai Muktevi

1. Introduction

After the mask mandates were set in place there were many instances where people refused to follow policy in several cases. The non-cooperation wasn't only regarding the masking policies but other policies as well such as getting vaccinated or social distancing. This was due to some political inclination where some leaders didn't support masking, people feeling uncomfortable with masking, or just outrageous conspiracy theories. Therefore, although we had mask mandates set in place there were still people that failed to follow through.

From the assessment in A4[1], We know when mask mandates were introduced in Harris County, TX and we also observed the trends in covid infection rates across April 2020 to October 2021. For this project I wanted to look into more qualitative data to understand how different kinds of people were influenced by various policies and how the level of their concern for the pandemic changed accordingly. Therefore, I took a look into Survey Data[2] regarding Harris County. From this survey I had data regarding employment status, education levels, demographics, measure of intent to follow certain basic policies, rating of coronavirus concern and even belief levels in myths and rumors about covid. All this information could be related to or even directly affect the way people understand the pandemic and mitigation policies. We learned in class about Thick Data[3] and how when we look at data regarding people there's a more *human centered aspect* to be considered than just looking at data points. My aim here was to take a more human centered approach to my analysis by looking at the story behind why some people were more concerned with the coronavirus versus others who seemed to be indifferent to it. I thought this Survey Data would be perfect to achieve my objective due to it's qualitative information on people during the pandemic.

The biggest policy that is believed to have had the most effect on curbing covid transmissions seemed to be the strict lockdown imposition and travel restrictions. These were some of the more extreme measures taken to reduce transmissions rates. Mobility struck me as an interesting metric to measure against covid transmission to see how different kinds of mobility within the county affected transmission rates. This is why I included Mobility Data[4] into my analysis.

My motivation for this is my compelling irritation as to why some people can't just follow the right covid protocols to keep themselves and others safe. From another perspective, I wanted to see what factors influenced people directly or indirectly to maintain covid protocols. I thought I could study in-depth to see if there was a more concrete data-driven answer to this question.

For people this study can be useful in multiple ways bringing clarity to the behavior of people during the pandemic. We can try to find what factors influence or relate to covid concerns and try to find ways to mitigate the risk of infection due to those factors. We can look at how more extreme policies, like restricted travel, actually affect transmission rates and create awareness to ensure more people follow such policies when advised in the future.

2. Background/Related Work

The coronavirus pandemic affected society in various ways and - on the flipside - different factors affected the way people responded to the coronavirus pandemic. My interests were piqued by how these two different sides interacted with each. The main question was, how did people react to the changes in society and livelihood. Did these changes affect the way people responded to more policies and restrictions during the pandemic? Did people have varying levels of concern for the coronavirus that was the result of these changes? These questions arose during my research regarding the pandemic and its effects on our country. In the following paragraphs I will explain how my research informed my hypotheses and analysis.

There are many cases where different aspects of people's livelihoods were affected in different ways. Unemployment was a big issue during this time. An article[5] I came across talks about how minorities were very severely affected by "pandemic-driven" unemployment. It talks of how "the state [of Texas] lost 1.4 million jobs from February to April 2020, and the unemployment rate shot up to 12.9 percent." They compare unemployment rates to that of the Great Recession. This made me want to look into unemployment rates in the county of Harris, TX, to see the employment status of people that were more concerned about coronavirus. I wanted to investigate to see if there was any relation between employment status and concern for the pandemic.

Similar to unemployment, another aspect to look at is education. I took education levels as a parameter as I believed that many were just ill-informed or not well educated to interpret the changes our society was going through. I believed that because of this they didn't respond in the right way to the required policies and mandates that were to be followed. A series of articles on outrageous behaviour across the country led me to this theory. For example, some of the comments from anti-maskers that I read in this article[6] seemed totally absurd. Some people didn't understand the seriousness of the issue and were indifferent to scientific claims. What's worse was that it became a political issue with

“President Donald Trump and many Republicans [having] spent months using them as a political lightning rod.” They weren’t supportive of the idea in the beginning making those that were politically inclined to one side biased in their views. Hence, political bias was also something I decided to study in this case. Regarding education, I thought it wasn’t very well informed and wrong of me to make the assumption that those with lower levels of education were less likely to follow through with policies after studying the issue of political bias further. So instead I wanted to test the opposite theory that if higher levels of education meant people were more inclined to follow through on mandates or if there was any relation at all between covid concerns and education levels.

Coronavirus restrictions that affected the most people were regarding travel. The mobility of people across various states and counties varied a lot during the pandemic. There are several papers discussing how COVID was transmitted across the world and on how travel increased the rate and spread of the virus across the world making it the pandemic that it is today. I, myself, participated in a hackathon organized by our MSDS program at the University of Washington studying the spread of COVID around the world and how various policies emerged to curb the rates of COVID in different countries[7]**.** This fueled my curiosity to look deeper into mobility. I wanted to observe this aspect of the pandemic as well and how it affected the number of COVID cases per day. For this reason I decided to explore mobility data.

My initial assumption was that those with more concern for COVID had more intentions to follow mandates and nation-wide policies and were more likely to do so. Informed by my research, I arrived at the following research questions and hypotheses:

Question 1

Is the Coronavirus Concern of a person influenced by their...

1. **Education levels,**
2. **Employment status,**
3. **Measure of intent to follow basic covid policies.**

Hypothesis: Coronavirus concern relates to Education Levels, Employment Status or Measure of Intent to following covid policies.

Question 2:

Do COVID cases per day show any meaningful relation with the various kinds of mobility data?

Hypothesis: COVID transmission changes with different kinds of Mobility in different ways.

3. Methodology

Due to the scope of this project, I didn't aim to answer all the questions I posed regarding my research but merely tried to understand the data that I had better. I wanted to see if some baseline assumptions that I made could be met and if I could test simple hypotheses based on these assumptions. In some areas I attempted to pursue my main research questions and others I merely followed through as the data guided my curiosity.

Survey Data Analysis [2][9][10]

First, I decided to look into Survey Data obtained from C3.ai[8]**.** My main motivation for looking into this dataset was due to it's qualitative nature. They also had a thorough and well built python API to access this data. Survey data is tough to work with but it is really more "Thick Data" compared to other kinds of data. This dataset delves down to the human level of each data point by describing people's outlook towards various aspects of the pandemic and its effects. Each person that takes the survey is able to directly convey their perspective on the pandemic by answering a survey. I took up the challenge of analysing Survey data although it's tricky to do so with basic statistical methods of analysis.

Collecting the data for Harris county, TX, wasn't possible as there wasn't a field mentioning the exact county of the participant. This could be due to privacy reasons which is of great importance as we've learned through this course. Therefore, I decided to proceed with State-level data for this analysis and filtered survey data based on participants from the state of Texas. Information on how these participants were sourced wasn't mentioned. They could be Turk workers but this is just my speculation. Further investigation regarding the actual source of the data led me to this github repository[10] which claimed that "the data provided in this repository is made available for public, non-commercial use by Swayable in partnership with TapResearch. The data includes symptom prevalence, demographic and political opinion data collected from thousands of respondents since April, 2020."

After collecting the data I began to explore the various fields and attributes of the data. There was a wealth of information regarding demographics, employment status and education levels. Survey questions posed were mostly on a scale of 0 to 100 or 0 to 10. There was also data on the measure of intent to follow different COVID mandates/policies, political inclinations and belief in myths as boolean entries. The attribute that stood out to me was regarding Coronavirus Concern measured on a scale of 0-10, 0 meaning “not at all” and 100 meaning “extremely concerned”. I found this interesting and looked at the distribution of concern among the participants. I also proceeded to look into the political inclination of the participants and especially their Trump approval ratings. I recorded my findings and proceeded to pursue my first hypothesis.

I centered my analysis around COVID concerns among the participants due to my initial assumption that those with more concern for COVID had more intentions to follow mandates and nation-wide policies and were more likely to do so. This was in order to find out how people’s concern for COVID was affected by various aspects of their lives and by societal changes.

My first hypothesis was inclined towards understanding the relationship between concern for COVID and employment, education, and intent to follow various policies. First, prior to this, as an extension of the A4 assignment I looked into COVID concern and intent to follow the masking policies. I pursued a more qualitative assessment of the mask mandates that we looked at from A4 as I believed we couldn’t get the entire story regarding how masking affected the rate of COVID transmissions without observing how people reacted to following the policy. This led me to finding the Survey data and further looking into all these other attributes.

I observed a scatter plot of just intent to follow masking policies and COVID concern. Then I checked Pearson's correlation between them. I didn’t find any interesting or meaningful revelations and decided to proceed with my analysis of COVID concern and education levels first. I looked at the distribution of education levels and found that most had finished high school and had some level of college education. I performed an ANOVA test to see if there was significant difference in the mean coronavirus concern between the different education levels then recorded my results. I did the same analysis for employment status. Although in this case I had to process the employment status as multiple statuses were mentioned as a comma-separated list to answer the question how it had changed for a participant since January 1, 2020. Naturally, I took the latest status as the most recent and performed the analysis. The employment data seemed to not meet the distribution criteria required for categories in data used for an ANOVA test. Therefore, I also performed a non-parametric Kruskal-Wallis test for the same hypothesis. I recorded my findings which I will elaborate more on in the next section.

Next I looked into the intent to follow basic policies like Masking, maintaining Six Feet distance, Staying at Home and Washing Hands. I wanted to understand how COVID concern related to the kind of policies that were followed. I wanted to establish the simple theory that people with more concern had more intent to follow these policies. I wanted to do this as most of my results so far didn't yield any interesting relationships so I decided to look at a simpler more obvious result regarding this Survey data if it followed a baseline understanding of the various attributes it came with. I performed a Multivariate Linear Regression with the coronavirus concern as the response variable and the intent to follow each of the basic policies as predictors. This would observe how much intent to follow each policy varied with the increase in coronavirus concern. I recorded my results and moved on to study the mobility data.

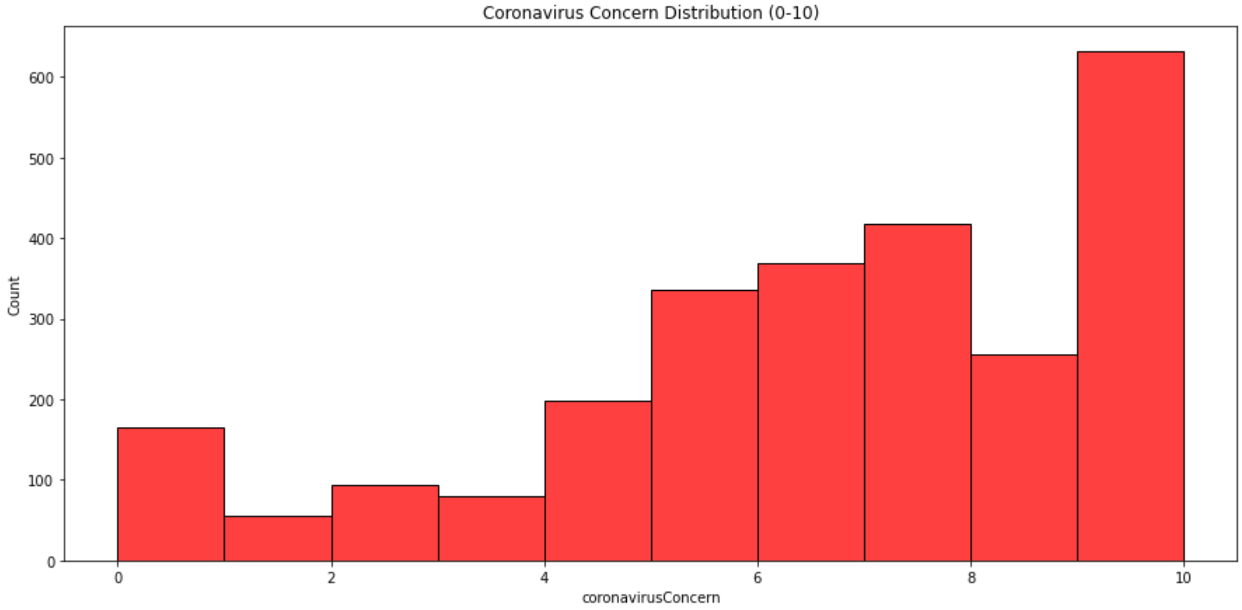
Mobility Data Analysis [4][9][11]

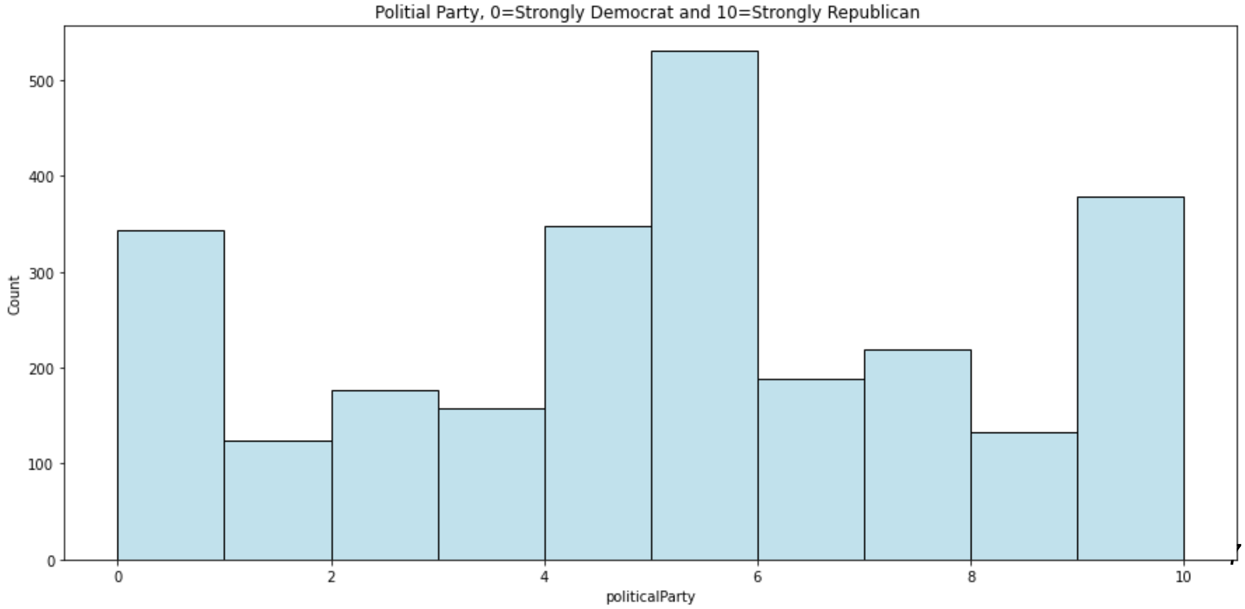
Mobility data also came from the CS.AI[4] source. I chose to observe mobility data as it would give us a better understanding of how people restricted their travel after lingering effects of the lockdown. Mobility data was sampled at the County level therefore we had data only on residents from Harris County, TX. There were two kinds of data that I had available, one collected from Google maps and the other from Apple maps. I was immediately concerned with privacy requirements and researched to find that this data was collected well within the accepted parameters of privacy guidelines for the good of understanding transmission rates better. There was one distinct feature about each data set that stood out. Google tracks data regarding mobility based on the "to" destination (parks, groceries, etc.) whereas Apple tracks the type of mobility used (walking, public transit, etc.). I performed a Multivariate Linear Regression to understand how much each type of mobility varied with the number of COVID cases per day. I realized that Google and Apple data couldn't be analyzed together and immediately implemented separate tests as they described two different aspects of mobility as mentioned. The test with Google data remained inconclusive but the test with Apple data was interesting which I will discuss in the next section.

4. Findings

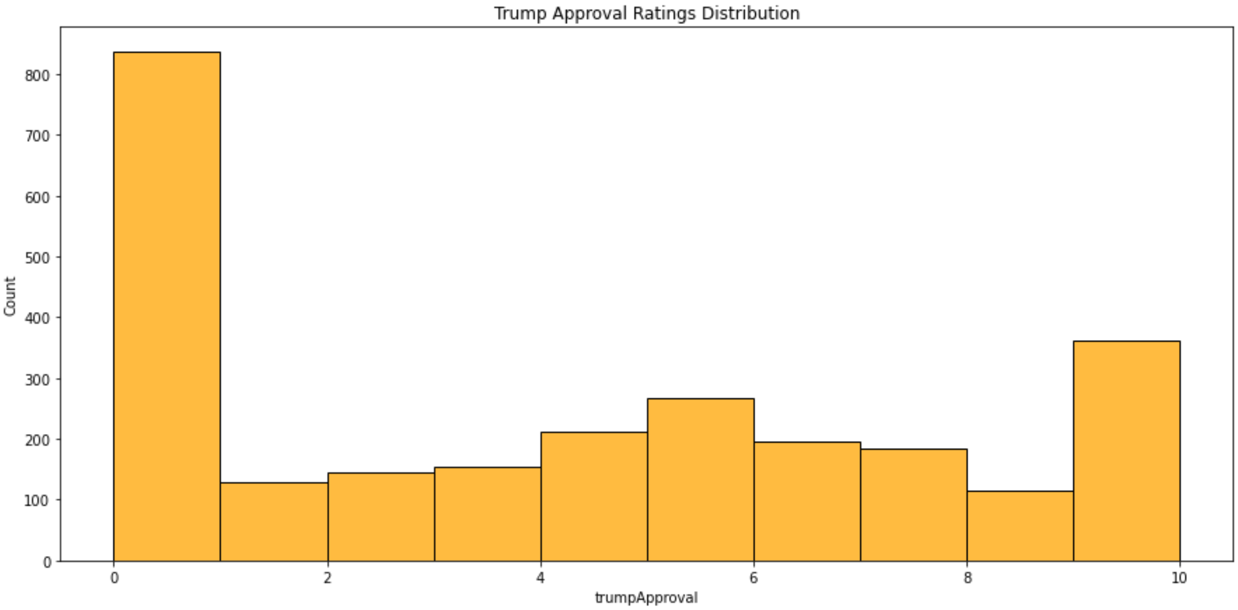
There were a couple of findings at the end of every analysis section that helped me decide on how to proceed forward with successive steps. We will discuss findings from the Survey Data analysis and then move on to the Mobility section.

On observing the coronavirus concern distribution in the histogram below we can see that it's fairly left-skewed showing that in this data set there are more people that have relatively higher ratings for concern. This is important because to measure against any other dimensions we should have enough data on coronavirus concern among people to see how it is affected by these other dimensions. Next let's look at the histogram regarding political bias to assess what political inclination this sample of people have from the



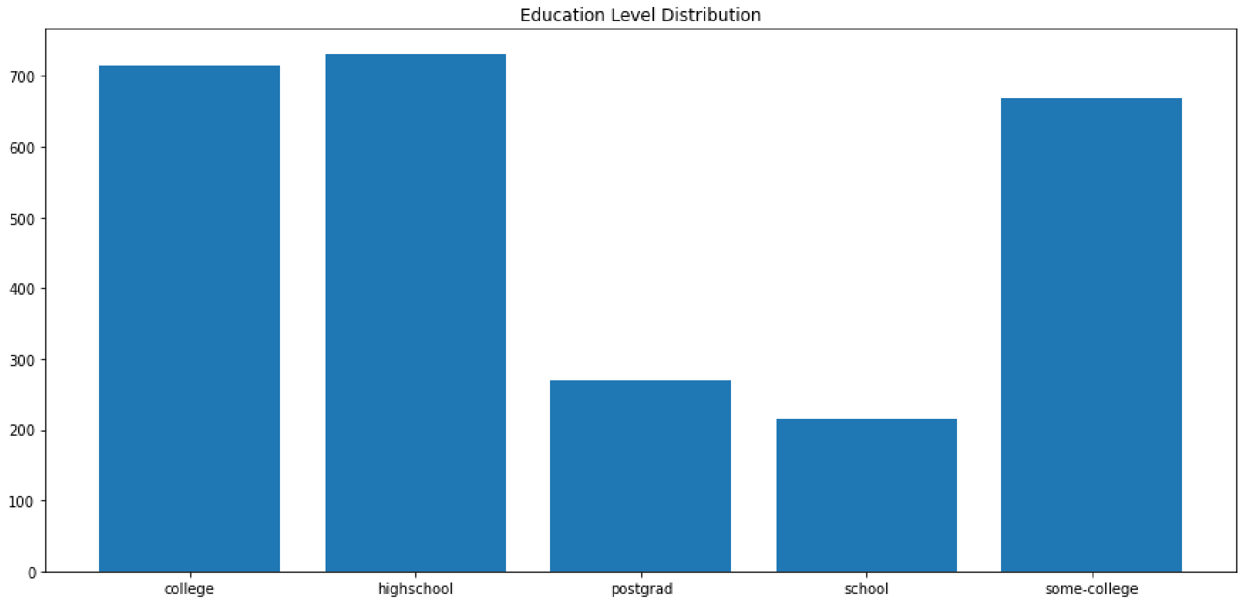


population. It looks like the political inclination is fairly distributed. Although, there is a bit more of a democratic majority. Since I was looking into Trump's early responses to CDC recommendations I decided to gauge the Trump approval ratings regarding COVID in the next graph. It was surprising to find many had a poor opinion of Trump in this context as I was expecting that people from Texas were mostly right leaning and trump activists. This also goes to show political inclination doesn't really align with Trump approval in this data

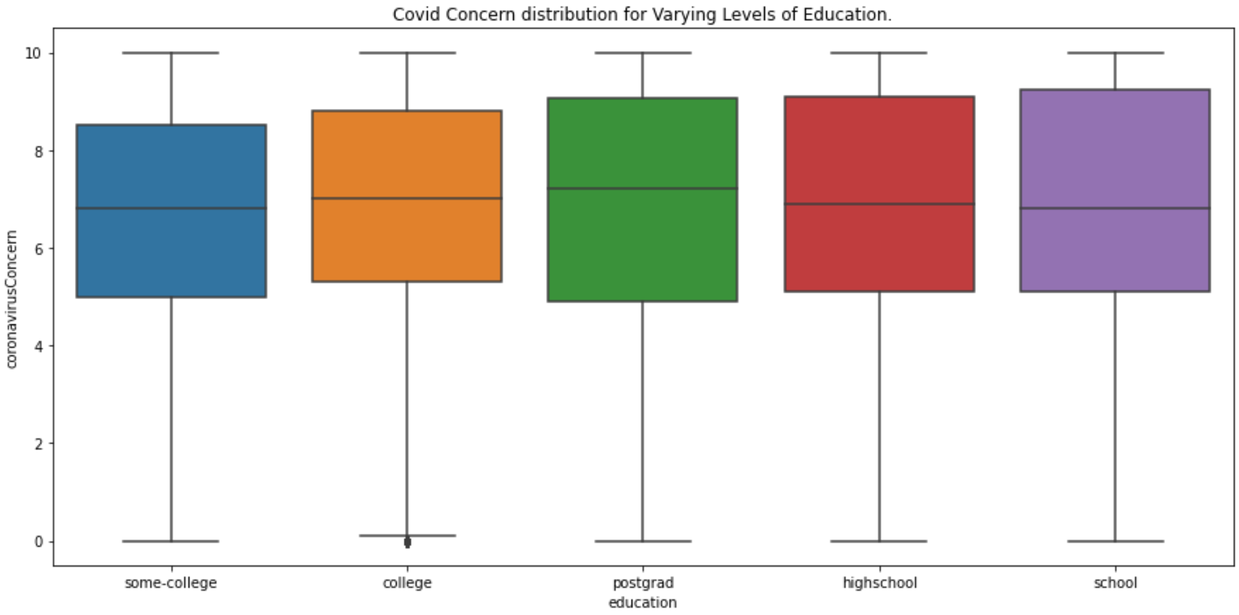


due to the distinct differences in distribution. This could mean that even strong republicans disapproved of the way Trump handled the pandemic. However, this is just a personal speculation and I might be biased. It's interesting to observe the low Trump Approval Ratings for this state - Texas - which is predominantly right. Political bias plays an important role in issues that can be made political like trusting certain government policies especially during the pandemic. Therefore, it's important to have this information to judge the political inclination of this crowd which may be the result of selective sampling or bias in the data.

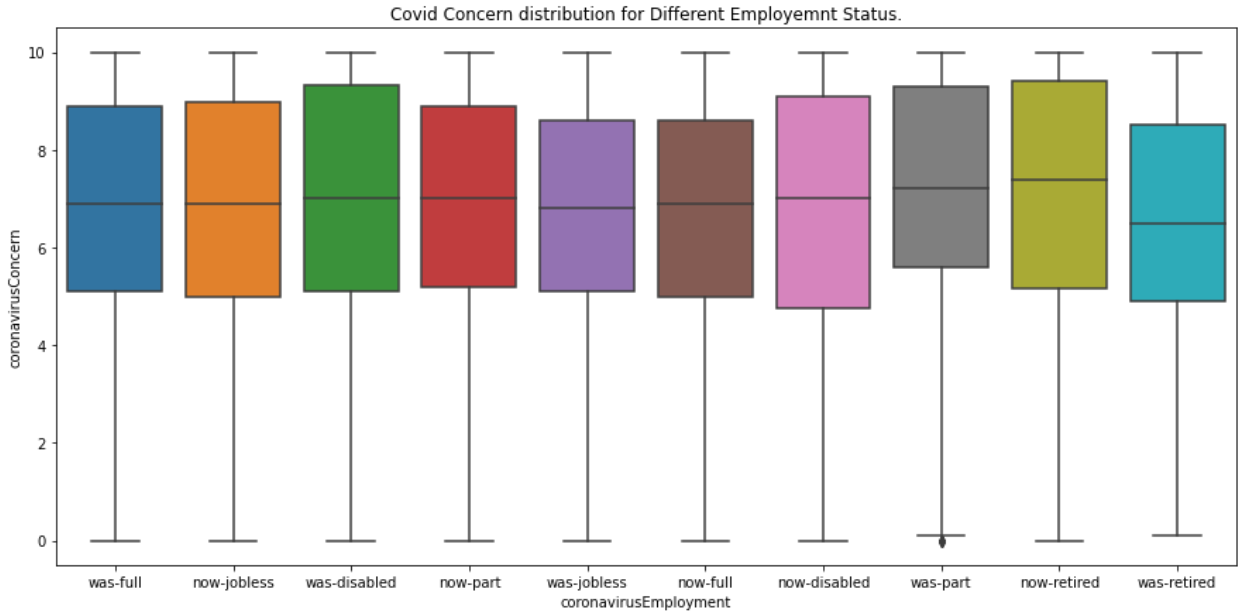
First let's analyse the investigation into the relation of education levels with coronavirus concern. Looking at the distribution in the following graph:



We can see that most participants had some level of college or high school education. Now we see how the distributions of concern compared across different education levels, we look at the following boxplot:

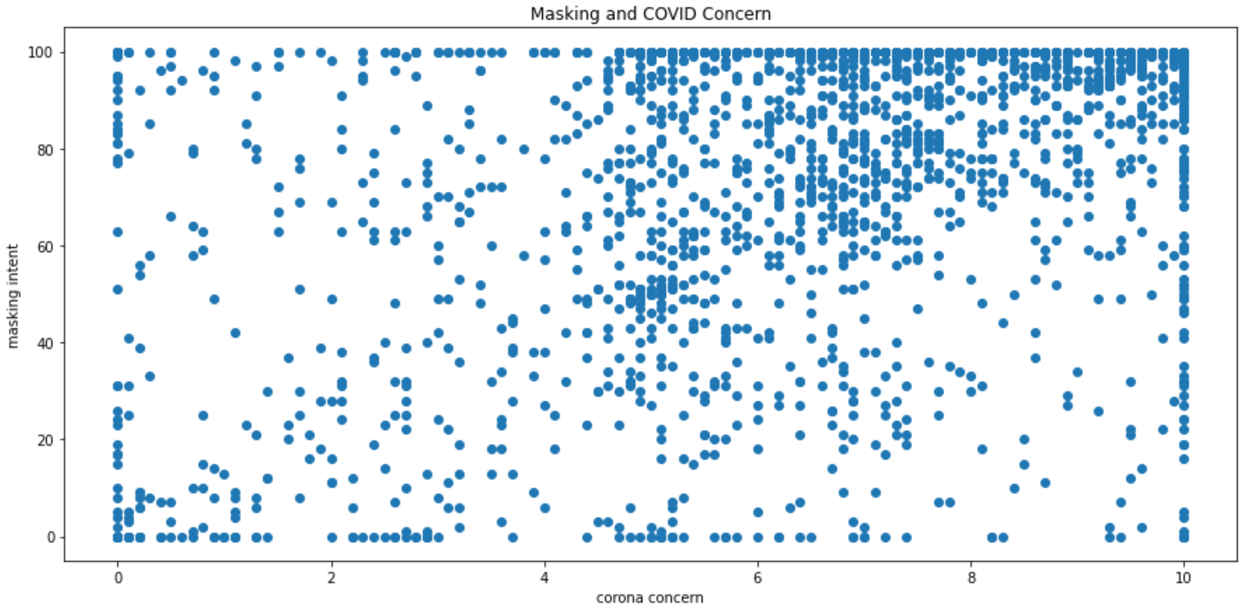


The distribution of coronavirus concern doesn't change by much with different education levels. The ANOVA test that I performed confirmed that the mean concern of each level of education doesn't show significant difference by p-value of 0.42. I performed a similar analysis for employment status and coronavirus concern. The high p-value shows that we can't reject the null hypothesis that there is no significant difference between the means of each category.

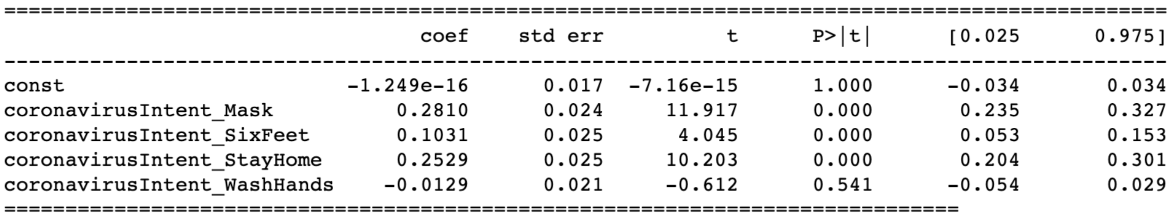


We observe that some distributions in case of employment status may seem a bit more skewed compared to the others. However, I performed an ANOVA to check if the difference in means was significant but arrived at the same result. The high p-value showed that we can't reject the null hypothesis that there is no significant difference between the means of each category. The Kruskal-Wallis test showed similar results with a p-value at 0.20.

This confirmed that based on the information from this dataset we can say that there was no evidence to support the claim that coronavirus concern among participants had any relation with employment status or education levels.



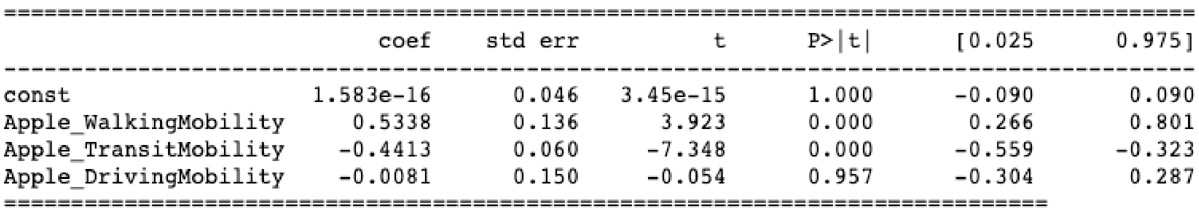
Moving on to other dimensions of the data, graphing a scatter plot of the coronavirus concern versus the intent to follow masking policies seemed haphazard at first but also showed a relation based on density of the points. There wasn't a clear linear relation but it looked like people with more intent to mask had higher concern regarding covid. I hypothesized that COVID concern would be directly related to intent to follow various policies. Use Multivariate Linear Regression to observe how intent to follow each policy affected COVID concerns among participants. I found that Masking, Six Feet and Stay at Home policy intent significantly affect the levels of COVID concern among these participants.



```
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                     -1.249e-16      0.017  -7.16e-15      1.000      -0.034       0.034
coronavirusIntent_Mask        0.2810      0.024     11.917      0.000       0.235       0.327
coronavirusIntent_SixFeet     0.1031      0.025      4.045      0.000       0.053       0.153
coronavirusIntent_StayHome    0.2529      0.025     10.203      0.000       0.204       0.301
coronavirusIntent_WashHands  -0.0129      0.021     -0.612      0.541      -0.054       0.029
==============================================================================
```

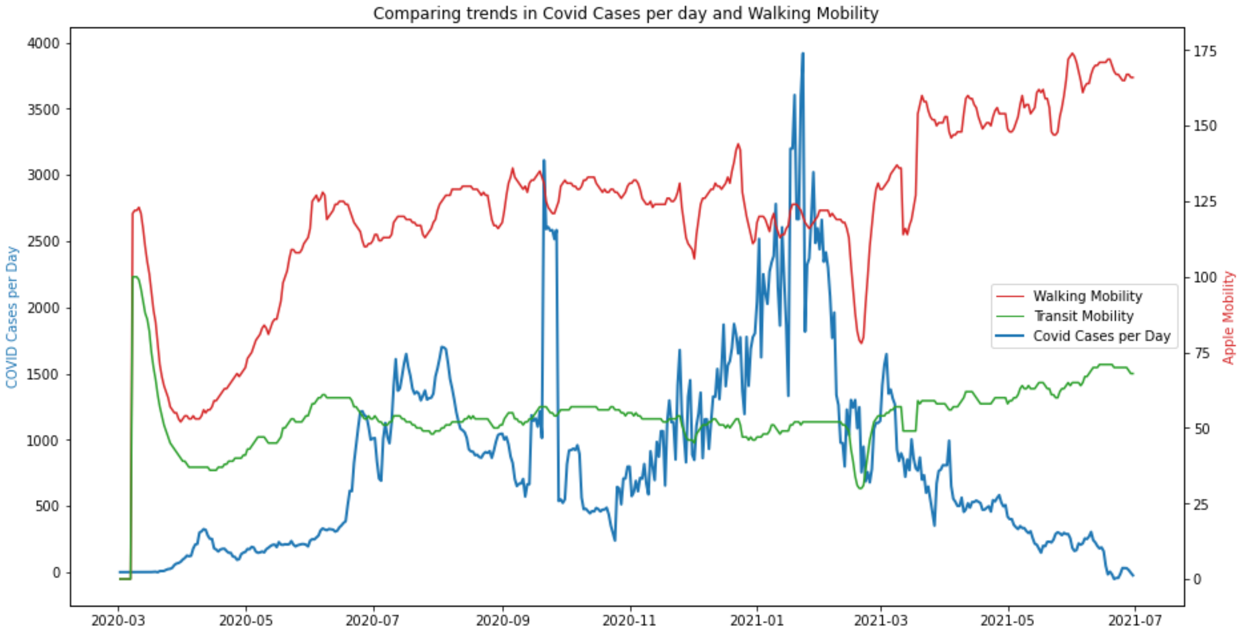
It looks like masking has a higher, more positive contribution to coronavirus concern. It's also interesting that data on intent to Washing Hands isn't significant enough based on its p-value to be contributing to concern among participants. It turns out that people with more concern will tend to follow more extreme policies like masking and staying at home.

Regarding mobility data, the regression analysis on different types of mobility from apple user data showed interesting results. Apple mobility data shows that the number of COVID cases per day has a significant direct relation with Walking route requests and an inverse relation with Transit route requests. Driving mobility data has too high of a p-value to make any significant inferences.

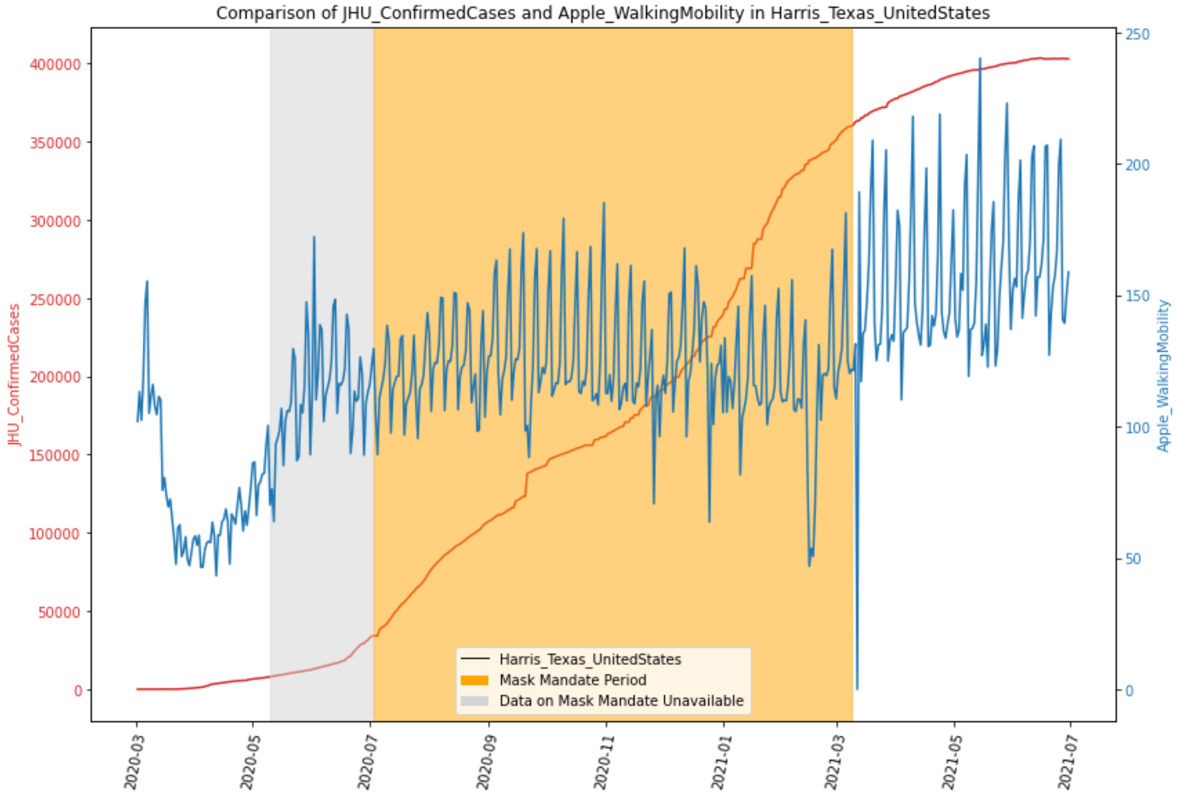


```
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  1.583e-16      0.046   3.45e-15      1.000      -0.090       0.090
Apple_WalkingMobility     0.5338      0.136      3.923      0.000       0.266       0.801
Apple_TransitMobility    -0.4413      0.060     -7.348      0.000      -0.559      -0.323
Apple_DrivingMobility    -0.0081      0.150     -0.054      0.957      -0.304       0.287
==============================================================================
```

There isn't any definitive interpretation of this result rather than speculation on how this relation could be possible and therefore its suggested that more research be conducted in this case before making any other conclusive statements about the relationship between the two.



For the sake of comparing the two types of mobility with cases per day in a visual representation I designed the graph above. It does show different trends in both cases but I believe there are other factors at play here and believe I don't have enough information to make any more definitive conclusions. More research is needed in this analysis.



I tried investigating more by making multiple graphs like the one above to further observe various factors together in a single visualization. It appears that mobility trends may have changed during mask mandates and with the change in COVID infection rates but I didn't test these hypotheses for sound statistical proofs. This visualization is also not the best way to observe this but I included this in for documentation.

5. Discussion

My findings regarding the survey data are important because they show that certain factors don't relate to what causes concern for the coronavirus pandemic. Data on these factors bring in a more human centered viewpoint to the people affected by the pandemic and allow us to understand the people behind the masks better. Survey data accounts for qualitative analysis where each data point has a story to tell about a person and their perspectives. We were able to establish that education and employment status don't have much to do with the concern for coronavirus among people. We also established the simple fact that there were strong positive correlations among people with more concern for the virus to follow through on basic mandates and policies recommended by the CDC. Further research can now be done to bridge the gap in understanding what factors influence coronavirus concern directly and causally. Using this information we can help create better awareness during a pandemic to increase concern so that people follow the necessary measures to curb transmissions.

It also looks like coronavirus concern isn't distributed according to political inclinations in this case so that shouldn't be a strong metric to judge who will be willing to follow more policies for our participants. Sometimes the media tends to sensationalize such incidents depending on how biased the news source is to make issues against policy a political issue. Often it could just be a lack of the right information that doesn't inform people on why we need these necessary precautions. A more generalized large-scale study should definitely be conducted on how political bias influences decisions to follow policies using qualitative assessments.

Mobility can be further researched to better interpret the findings regarding walking and transit mobility and how they relate to the number of COVID cases per day. I believe there are definitely other factors at play here that I haven't taken into account and need to do a more in-depth analysis for better and more conclusive results.

6. Issues, Limitations and Implications

Initially the data was filtered in the default API config with 'coronavirusIntent_Mask' >= 75 which I later corrected to not include that filter. I performed my analysis again including the new data points and found mostly similar results. Hence, instead of 1,483 data points I now have 2,598 data points instead. The results I show now may not be the same values as presented but agree with the same findings.

The statistical analysis conducted in this project uses very fundamental tests that require a set of necessary assumptions to be met. With larger datasets it's easy to be lax about some of these assumptions but the smaller the data the more careful we have to be regarding meeting these assumptions. For example, the employment status data didn't meet the distribution and size requirements for each category so I proceeded with a Kruskal-Wallis test which is said to be a non-parametric form of ANOVA.

Survey data is usually tricky to deal with, especially regarding finding anomalous entries. I handled missing data by removing these data points all together with the assumption that the survey wasn't attempted earnestly in full completion. This isn't an accurate assumption. Different people have different opinions on how to perceive the intuition behind a scale in surveys. This also leads to noise in the data.

Survey data was small in size and was collected at the state-level. There's less granularity because it's not county-level information as was requested for this assignment. The small size makes it less generalizable. Imbalance across categorical attributes in the dataset will be misrepresentative of the population if the sampling was not done correctly. There isn't much information on how the survey data was collected or conducted so it's tough to understand the sampling techniques used. This could result in potential bias in the data. Speaking of bias, mobility data is also biased towards Apple maps users.

The conclusions drawn from these statistical analyses may need to be rechecked. There were no power calculations done to assess the strength of the tests in any case. It's not easy to make any causal inferences within the scope of this current analysis approach. Only speculative correlations and reasonings were discussed but further research needs to be conducted.

The biggest implication arises due to the presence of ordinal data. The data regarding intent to follow policies and coronavirus concern levels are ordinal data values picked by human participants. Ordinal data is tricky to deal with using statistical methods and I am not aware of all the appropriate statistical methods used for this kind of research.

This is by no means a perfect analysis and still has many flaws and issues to be discussed. There can be many improvements made here to get better results. However, this establishes a useful baseline to be regarded as a preliminary analysis to foster further research.

7. Conclusion

Research Question 1

Is the Coronavirus Concern of a person influenced by their...

1. Education levels,
2. Employment status,
3. Measure of intent to follow basic covid policies.

Hypothesis: Coronavirus concern relates to Education Levels, Employment Status or Measure of Intent to following covid policies.

This establishes that qualitative survey data that looks at people as not just data points has more to tell about the human perspective of understanding the COVID pandemic and how it changed our lives and mindsets. This type of data describes each participant and their perspectives of the pandemic summarizing all various dimensions of their lives affected in different ways. Using this data to find out relations between each of these dimensions gives us insight into how coronavirus concern is shaped among people from Texas. From our analysis, we found that there is no evidence that COVID concern is influenced by or relates to the education level or employment status of the participants. We also established with quantitative analysis that people more concerned with coronavirus tend to follow through on basic policies recommended by the CDC thereby acknowledging the legitimacy of this survey data.

Research Question 2

> Do COVID cases per day show any meaningful relation with the various kinds of mobility data?
>
> *Hypothesis: COVID transmission changes with different kinds of Mobility in different ways.*

In the second case, we were able to show that the number of COVID cases per day has a significant direct relation with Walking route requests and an inverse relation with Transit route requests.

8. References

[1] Muktevi, V. S. (2021). A4-Common Analysis. Google Docs. https://docs.google.com/document/d/1xezA55-xJ3BlJEYi_nRL5p__cUEiq12qz1ZGqhWmIbg/edit?usp=sharing

[2] C3.ai COVID-19 API Documentation. (2021). Retrieved 13 December 2021, from https://c3.ai/covid-19-api-documentation/#tag/SurveyData

[3] Wang, T. (2016, December 5). Why big data needs thick data. Medium. Retrieved December 14, 2021, from https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7.

[4] Mike Ralphson . (n.d.). C3.ai COVID-19 API Documentation. C3.Ai. Retrieved December 13, 2021, from https://c3.ai/covid-19-api-documentation/#tag/OutbreakLocation/paths/%7E1api%7E11%7E1outbreaklocation%7E1evalmetrics/post

[5] *Spotlight: Pandemic Pushes Texas Minority Unemployment Beyond Highs Reached During Great Recession*. (n.d.). Dallasfed.Org. https://www.dallasfed.org/research/swe/2021/swe2101/swe2101e.aspx

[6] Stewart, E. (2020). *Anti-mask protesters explain why they refuse to cover their faces during the Covid-19 pandemic*. Vox. https://www.vox.com/the-goods/2020/8/7/21357400/anti-mask-protest-rallies-donald-trump-covid-19

[7] Muktevi, V. S. (2020). *Covid Analysis*. UW COVID Hackathon. https://uw-covid-hackathon.github.io/covid-visualization/

[8] C3.ai. (2021, September 1). C3 AI - Enterprise AI. C3 AI. https://c3.ai/

[9] C3.ai. (2021b, September 27). *Creating a Unified COVID-19 Global Resource in Record Time*. C3 AI. https://c3.ai/customers/covid-19-data-lake/

[10] *GitHub - swayable/covid-19-data: COVID-19 related data collected since April*, 2020 by Swayable. (n.d.). GitHub. https://github.com/swayable/covid-19-data

[11] *COVID-19 - Mobility Trends Reports*. (n.d.). Apple. https://covid19.apple.com/mobility

9. Data Sources

Survey Data:

- C3.ai COVID-19 API Documentation
- Creating a Unified COVID-19 Global Resource in Record Time - C3 AI
- swayable/covid-19-data: COVID-19 related data collected since April, 2020 by Swayable.

Mobility Data:

- C3 AI COVID-19 Data Lake
- Creating a Unified COVID-19 Global Resource in Record Time - C3 AI
- COVID-19 - Mobility Trends Reports - Apple

Thank You