How State and Inter-County Travel Affected People during COVID-19

Elizabeth Hollis/Mendoza, Shane May, Edgar Nolasco, Piper Varney

Indiana University

Data Mining

Team

Team members include Shane May, Elizabeth Hollis/Mendoza, Edgar Nolasco, and Piper Varney. Shane has skills in writing research papers, designing PowerPoint slides, and generally making papers and ideas more professional. They also have experience with R and RStudio, which will benefit the team when working with the data in R. Elizabeth has experience working in laboratories and pursuing a degree in Biology, which could come in handy when dealing with scientific terms concerning COVID-19. Edgar has worked with R and RStudio, and has professional experience working with data manipulation in their job. Piper has experience working with Excel, creating datasets and data visualizations which will help the team when depicting the data.

Introduction

The main idea of this project will be based on the Covid-19 pandemic, with a primary focus on how it traveled around this area (state and county travel). Being able to track how COVID traveled and response times to the pandemic is key, travel of COVID-19 is different in other places depending on population, while responses to vaccines can be based on alignment to religion or political alignment based on the area. Going off of the previous statement, people can also be a “wildcard” when it comes to the spread of a virus and public perception and behaviors can be influenced by culture, this could also be used with predictive classification such as mask wearing, politics, public views on vaccines and herd immunity, vaccine distribution rates and other factors can contribute to the virus being spread over shores and borders.

Our research can have many different actionable uses, provide insight on public health trends, and give patterns of social activity, among many other possibilities. We can focus on advances in coronavirus identification, tracing methods, and threat level based on demographic, geographic, and other relevant data.

This helps in future pandemic/epidemics when comparing what responses should be appropriate for the disease that has come up, i.e. infectivity rates, mortality, etc. and deciding the right responses. As a major historical event during our lifetimes, Covid-19 is something that is still being discussed and discovered. Being a part of that discovery through this data mining project will intrigue us while also pushing us to discover new information and/or draw new conclusions regarding Covid-19.

With all this information in mind, we decided to focus on this research question:

Were university counties at a higher risk of COVID-19 due to the population and how that might affect IU colleges?

Literature Review

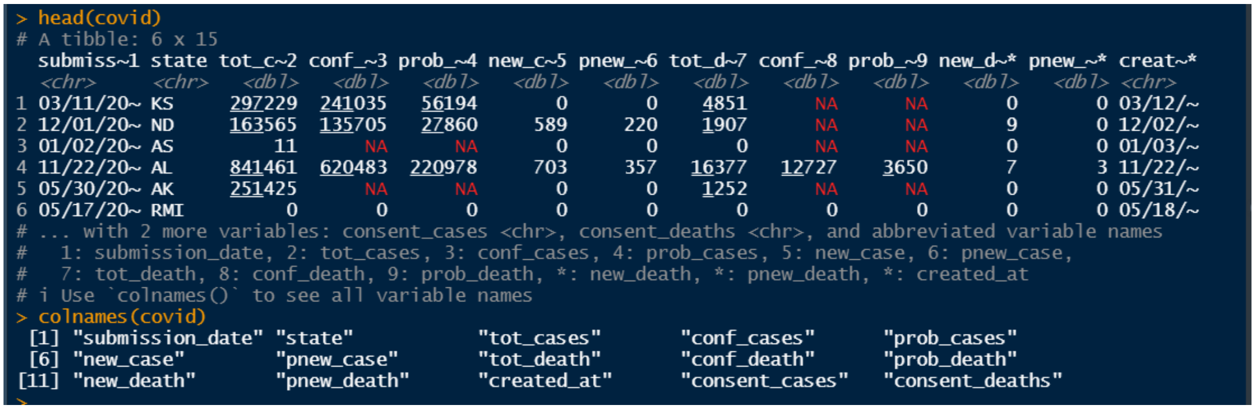
The field we are working in has been well established, since SARS COV-2 or COVID-19 became a world-wide problem in 2020, research has been conducted endlessly on the issue. There are many different research aspects to COVID-19, from research over vaccines and the virus's structure to response time and how COVID-19 has affected other industries.

During the height of the pandemic, restrictions and policies were put in place to help prevent the spread of COVID-19. Some common restrictions included stay-at-home orders, mask mandates, closings of schools and workplaces, etc. The Oxford COVID-19 Government Response Tracker (OxCGRT) is used to measure how the government (whether it be nation or state-wide) responded to the pandemic. In this study (Hale et al., 2020), the authors presented measures and variations across the US states.

During the entire pandemic travel and how the virus was spreading in different populations was a key in early research. Some studies used cell-phone pinging to using university schedules to track how COVID-19 spread and the time frames there were more cases. In one study (Yilmazkuday, 2020), researchers studied inter-county travel within the U.S. and the implications on new cases and deaths. They used a time frame of covering the period between January 21th, 2020 and September 2nd, 2020, in which they used cell-phone pinging and U.S. daily counters. In this study it was suggested that COVID-19 cases and deaths were lower in counties where people hadn't traveled across other counties as much as other counties in the quarantine time frame at the time, 14 days.

Data Resources

The following dataset titled, "United States COVID-19 Cases and Deaths by State over time" from data.cdc.gov, is real-time weekly COVID-19 statistics coming live from aggregated state and federal government databases, produced by the CDC and other participating health organizations (NCIRD(1), 2022). The data we were able to obtain from the CDC, as of October 20, 2022, stopped receiving weekly updates, so relatively recently, but it's still relevant to show the spread of the virus within the United States. The dataset contains data on every state, including major cities such as NYC and Chicago, and includes attributes in columns such as Total Deaths, Confirmed Deaths, New Deaths, New Probable Deaths, Total Cases, New Cases, Confirmed Cases, New Probable Cases, and other metadata including timestamps and sources. Here's a brief glance at the original dataset in R Studio prior to preprocessing:



```
> head(covid)
# A tibble: 6 x 15
  submiss~1 state tot_c~2 conf_~3 prob_~4 new_c~5 pnew_~6 tot_d~7 conf_~8 prob_~9 new_d~* pnew_~* creat~*
  <chr>     <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <chr>
1 03/11/20~ KS     297229  241035   56194       0       0    4851      NA      NA       0       0 03/12/~
2 12/01/20~ ND     163565  135705   27860     589     220    1907      NA      NA       9       0 12/02/~
3 01/02/20~ AS         11      NA      NA       0       0       0      NA      NA       0       0 01/03/~
4 11/22/20~ AL     841461  620483  220978     703     357   16377   12727    3650       7       3 11/22/~
5 05/30/20~ AK     251425      NA      NA       0       0    1252      NA      NA       0       0 05/31/~
6 05/17/20~ RMI         0       0       0       0       0       0       0       0       0       0 05/18/~
# ... with 2 more variables: consent_cases <chr>, consent_deaths <chr>, and abbreviated variable names
#   1: submission_date, 2: tot_cases, 3: conf_cases, 4: prob_cases, 5: new_case, 6: pnew_case,
#   7: tot_death, 8: conf_death, 9: prob_death, *: new_death, *: pnew_death, *: created_at
# i Use `colnames()` to see all variable names
> colnames(covid)
 [1] "submission_date" "state"           "tot_cases"       "conf_cases"      "prob_cases"
 [6] "new_case"        "pnew_case"       "tot_death"       "conf_death"      "prob_death"
[11] "new_death"       "pnew_death"      "created_at"      "consent_cases"   "consent_deaths"
>
```

Referenced in the Literature Review part were studies being done over inter-county and state travel using cell-phone pinging and looking at university schedules. Since data could not be recovered for the cell-phone pings, as it was done by the researchers themselves in another paper, we decided to create our own dataset for both inter-county and state travel. To create our own datasets we used the CDC COVID-19 tracker on the state level, tracking all the states around Indiana including Indiana, Kentucky, Illinois, Ohio, and Michigan. Time periods were chosen corresponding to periods within the last two years showcasing many sides of the pandemic, such as the start, when cases were down, holiday seasons (typically when people come together and could be a key for transmission), new mutations, etc. For the inter-county dataset, we chose only a few counties such as the counties that host the Indiana University campuses, playing off one of the studies that looked at transmission of COVID-19 in college students which are the prominent population to have no symptoms and thus spread the virus easily. The counties studied were Monroe, Marion, Henry, Howard, Lake, St. Joseph, and Floyd. The time periods were the same as the State dataset.

Description of the Data

From the CDC dataset, we can make the inference that the mean number of total cases which is much larger than the median was at the beginning of COVID-19 during the initial spread of the virus during a time when the vaccine was either not developed yet or hadn't been equally distributed and administered, as well as other circumstances revolving around each state like population, lockdown measures, etc.

The inter-county and state dataset we created contains both qualitative and quantitative data. In regards to the COVID-19 research question and spread in surrounding areas, these datasets will help to show how on a state level the surrounding areas compare to Indiana - whether there is a distinct difference in cases during the time periods we look at and on a county level of the university campuses to see in-depth how COVID-19 spread in our direct communities. The datasets will mainly focus on the cases and the time periods as this will showcase our understanding of how COVID-19 transmitted during the last two years.

	Descriptives and Statistical data				
State	**Indiana**	**Kentucky**	**Ohio**	**Illinois**	**Michigan**
Mean	122,351.125	97,144.75	180,467.4	216,436.1	136,228.9
Median	119596	81168	132267	140973.5	87772.5
Mode	N/A	N/A	N/A	N/A	N/A
SD	1101108.3	96690.33	187197	238696.9	15757136
Minimum	479	194	705	1862	2465
Maximum	339888	306353	565141	748137	493321
Count	8	8	8	8	8

Table 1. State descriptives and statistical data table - shows the different states surrounding Indiana and how COVID-19 cases compare to each state.

	Descriptives and Statistical data						
County	**Monroe**	**Marion**	**Wayne**	**Howard**	**Lake**	**St. Joseph**	**Floyd**
Mean	1272	16248	1466	1915	6761	4308	1401
Median	1344.5	12939.5	1333.5	1677.5	5061.5	3244.5	1330
Mode	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SD	858.5	15502.21	1396.73	1774.78	5645.83	4104.72	1377.01
Minimum	41	1078	9	39	363	123	74
Maximum	2402	48868	4423	5647	14416	12022	4432
Count	8	8	8	8	8	8	8

Table 2. County descriptives and statistical data table - shows the different counties in Indiana which host the Indiana University campuses and how COVID-19 cases compare to each county/IU campus.

Hypothesis and Goals

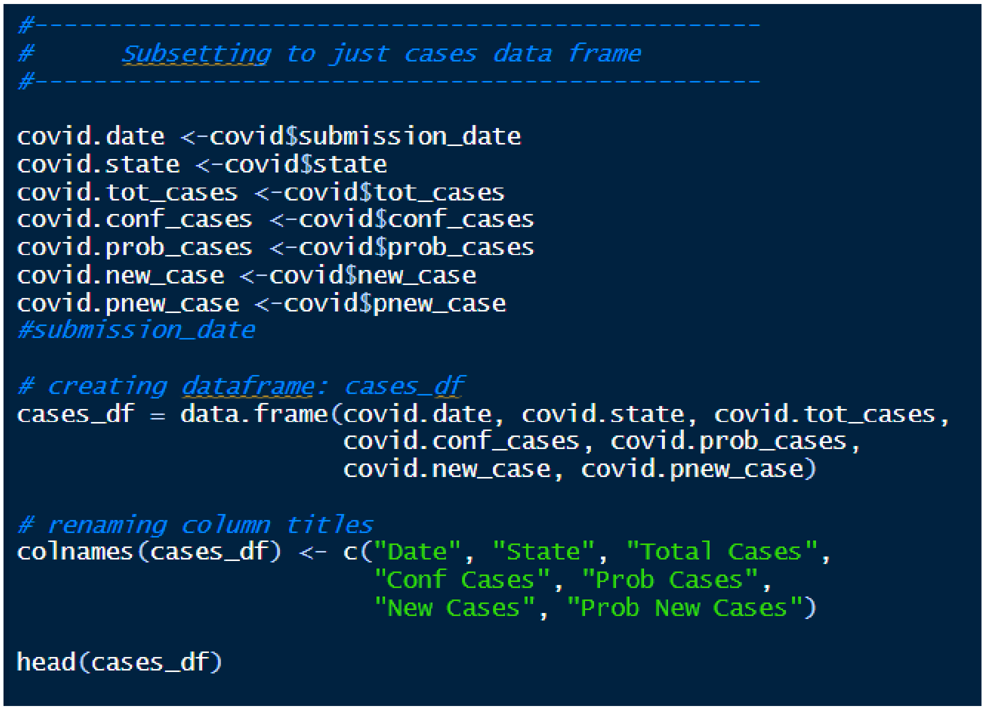
One inference or analysis we would like to make is how spread of COVID-19 is like in different states as well as counties in Indiana and why. Since we followed the states surrounding Indiana, we can see how these all compare to each other - we may see differences in states based on whether their political leanings are red or blue, or we may see differences based on population demographics, such as age, sex, and race. Both datasets follow a specific time period, with most of the periods being around the holidays, beginning of school, summer break, etc. This was designed since we believed this is when most cases would be as more people are traveling or come from traveling and could potentially spread COVID-19 to others.

For all dataset biases, we have the expectation that states with bigger populations will have higher numbers of cases and deaths, counties with more enrolled students in their universities will have higher cases than the other IU universities, and states that lean blue will have higher numbers of vaccines administered while red states have a lower number of administered vaccines. We also already know that age plays a factor so communities we are familiar with, in terms of counties in Indiana we know, we could have a bias about if we know whether the population has a majority of older people or not.

The end goal of our project is primarily to help people and contribute to the ongoing COVID pandemic response. We want to find correlations in the state and inter-county datasets between the different states and counties and why these correlations exist such as giving explanations.

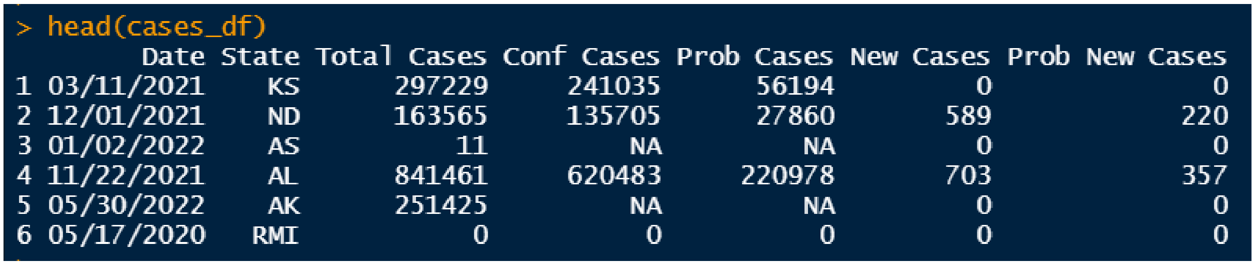
Research Design

Data cleaning and preprocessing is a significant step in the data mining process, sometimes 90% of work building a data model can be cleaning the data. For this we've used several methods, including basic filtering and subsetting, for example:



```
#------------------------------------------------
#      Subsetting to just cases data frame
#------------------------------------------------

covid.date <-covid$submission_date
covid.state <-covid$state
covid.tot_cases <-covid$tot_cases
covid.conf_cases <-covid$conf_cases
covid.prob_cases <-covid$prob_cases
covid.new_case <-covid$new_case
covid.pnew_case <-covid$pnew_case
#submission_date

# creating dataframe: cases_df
cases_df = data.frame(covid.date, covid.state, covid.tot_cases,
                      covid.conf_cases, covid.prob_cases,
                      covid.new_case, covid.pnew_case)

# renaming column titles
colnames(cases_df) <- c("Date", "State", "Total Cases",
                        "Conf Cases", "Prob Cases",
                        "New Cases", "Prob New Cases")

head(cases_df)
```

As we've described above, we've decided to subset and clean the data to be more clear, so we split the COVID-19 dataset into two, one for cases and one for deaths, and excluded some of the other metadata. We will focusing on the cases dataset and information we gain from that specifically though.



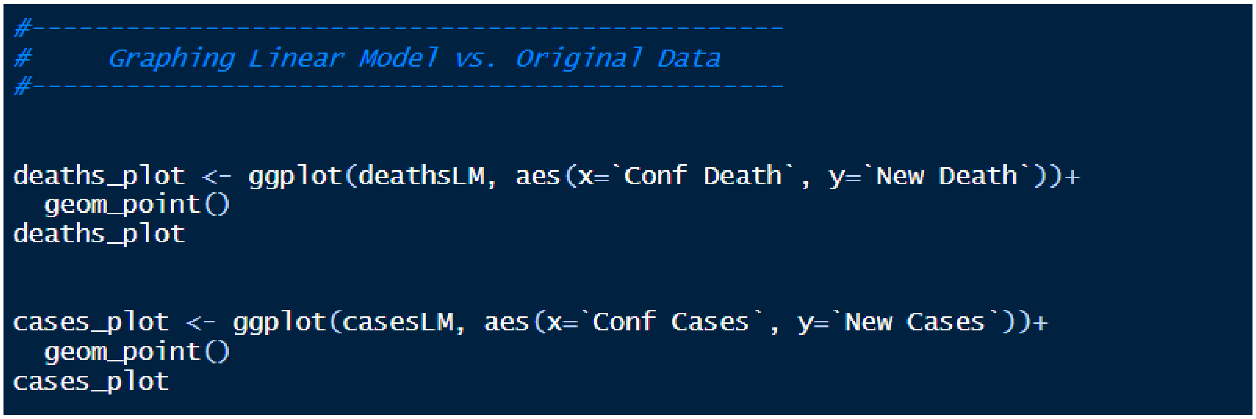
```
> head(cases_df)
        Date State Total Cases Conf Cases Prob Cases New Cases Prob New Cases
1 03/11/2021    KS      297229     241035      56194         0              0
2 12/01/2021    ND      163565     135705      27860       589            220
3 01/02/2022    AS          11         NA         NA         0              0
4 11/22/2021    AL      841461     620483     220978       703            357
5 05/30/2022    AK      251425         NA         NA         0              0
6 05/17/2020   RMI           0          0          0         0              0
```

The new subset above has renamed labels to have more clarity, such as changing "Conf_Cases" to "Confirmed cases." The original dataset had some ambiguous labels and hard-to-read abbreviations, such as "pnew" meaning "Probable New." In addition we need to clean the observations to only complete observations with no missing null values, which R Studio automatically shows as "NA."

Since we created our own dataset, it has already been cleaned and segmented when creating the dataset initially. We will focus more on the monthly cases, the month/year that these cases occurred, and state/county while comparing them. The programming language R will be used to analyze, visualize, and form conclusions on all the datasets.

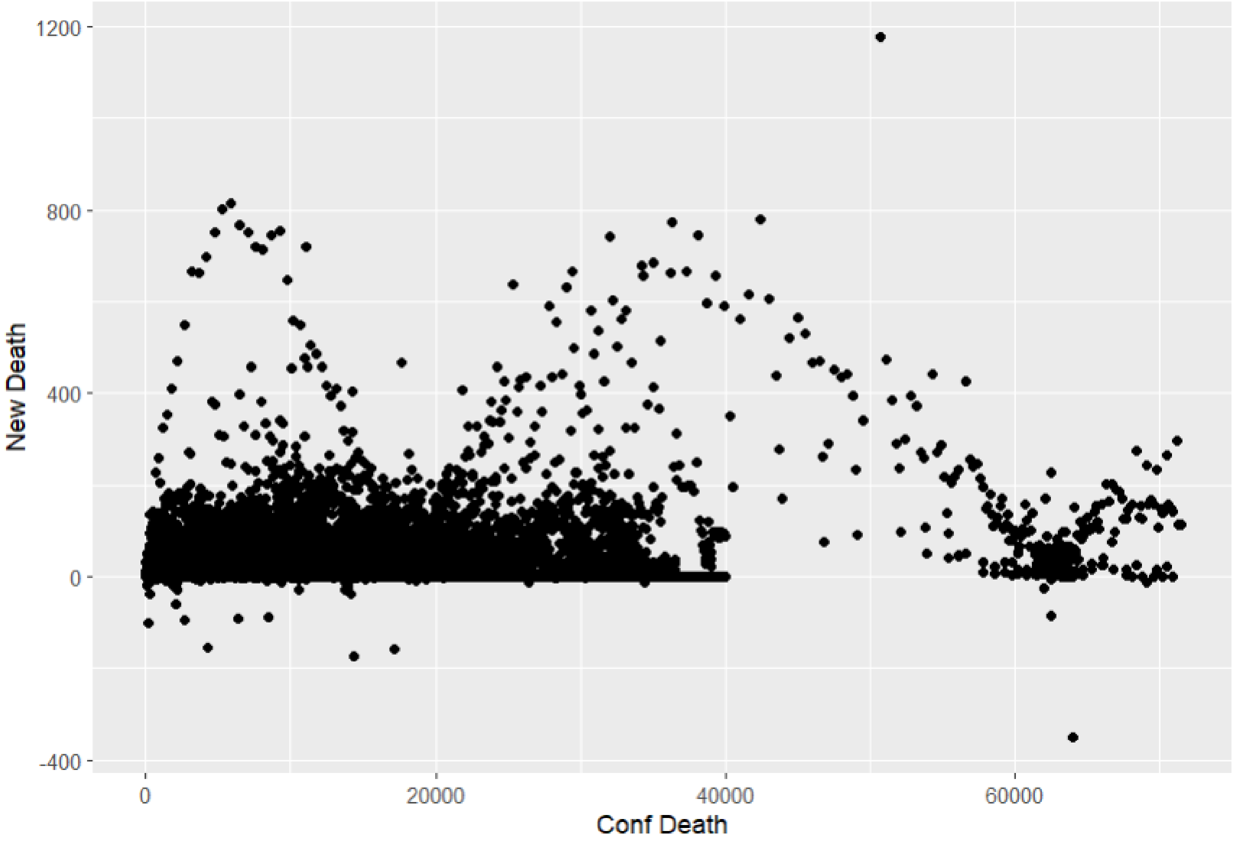
Methods, Visualizations, and Analysis

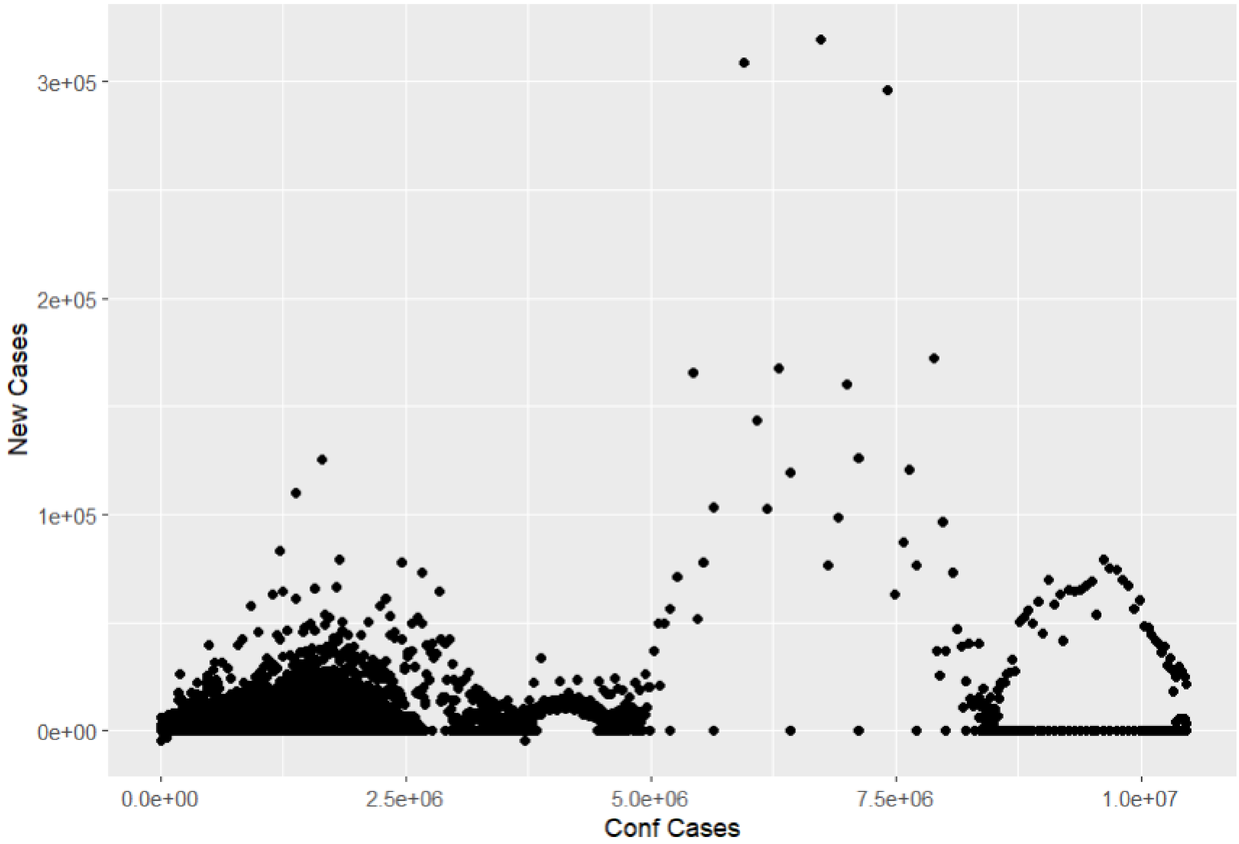
Using data and cleaning provided in previous sections, the CDC state cases and deaths datasets were simplified with plotting and linear regression:



```
#-------------------------------------------------
#      Graphing Linear Model vs. Original Data
#-------------------------------------------------


deaths_plot <- ggplot(deathsLM, aes(x=`Conf Death`, y=`New Death`))+
  geom_point()
deaths_plot


cases_plot <- ggplot(casesLM, aes(x=`Conf Cases`, y=`New Cases`))+
  geom_point()
cases_plot
```

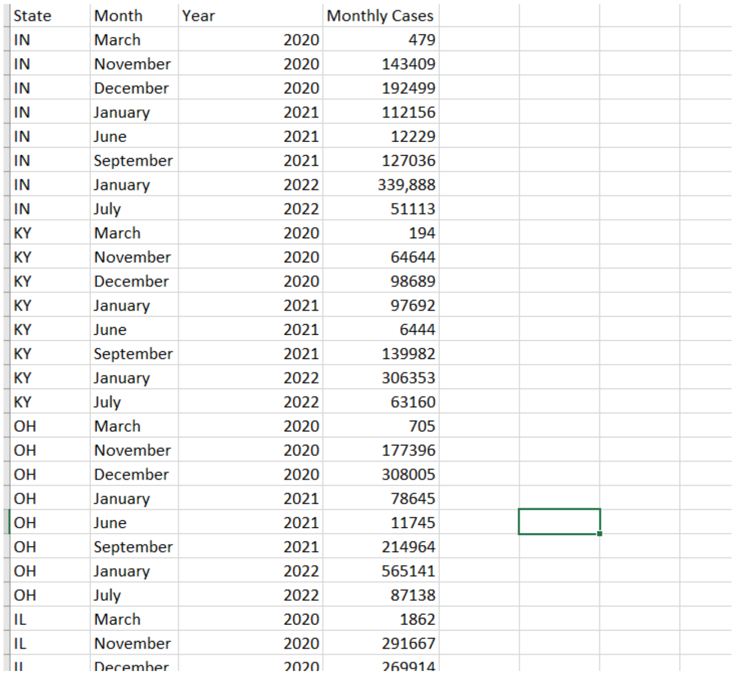
A visualization of the plotting shows as this:





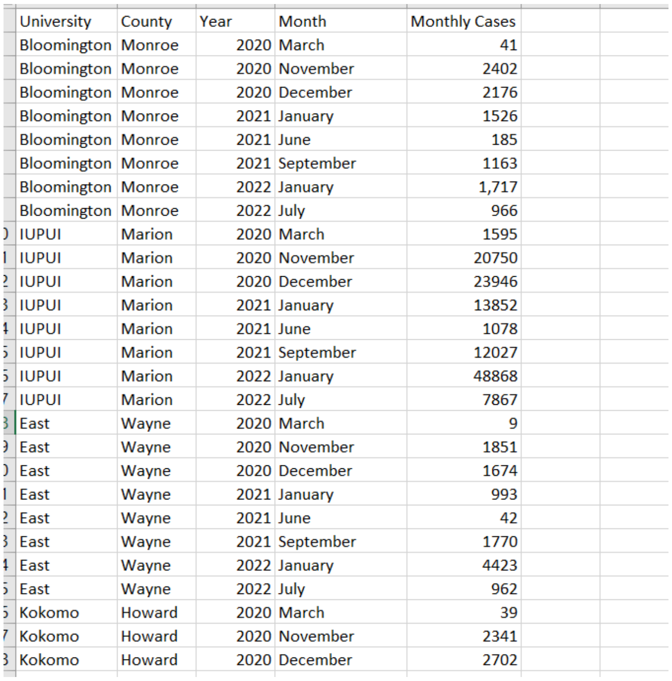
‘New Cases’ represent the number of cases that were new cases to someone who had not already been infected with COVID-19, and ‘Confirmed Cases’ is the total of officially observed cases, and obviously ‘New Deaths’ and ‘Confirmed Deaths’ are the number of new and confirmed deaths officially observed and reported. Both of these geometric plot graphs display similar waves and patterns of COVID-19, following the same curve, and both likewise show the same instances of outliers. While the graph of deaths looks as if there’s more observations, the X and Y-axis show cases to far surpass the number of deaths, which is positive and means many people were infected with COVID-19 and only a small percentage died from it. We can also see large concentrations and then sparse data points in the same time frame, representing large outbreaks versus smaller, isolated outbreaks, and this can be easily cross-referenced with locations that lifted COVID-19 restrictions, especially travel restrictions, ending in many more people traveling and infecting others across borders.

Since the state and county datasets were made by our group, they were done as simply as possible. Here are a few visualizations in Excel to showcase off the datasets:



State	Month	Year	Monthly Cases
IN	March	2020	479
IN	November	2020	143409
IN	December	2020	192499
IN	January	2021	112156
IN	June	2021	12229
IN	September	2021	127036
IN	January	2022	339,888
IN	July	2022	51113
KY	March	2020	194
KY	November	2020	64644
KY	December	2020	98689
KY	January	2021	97692
KY	June	2021	6444
KY	September	2021	139982
KY	January	2022	306353
KY	July	2022	63160
OH	March	2020	705
OH	November	2020	177396
OH	December	2020	308005
OH	January	2021	78645
OH	June	2021	11745
OH	September	2021	214964
OH	January	2022	565141
OH	July	2022	87138
IL	March	2020	1862
IL	November	2020	291667
IL	December	2020	269914

This is the Excel dataset for the state cases by month

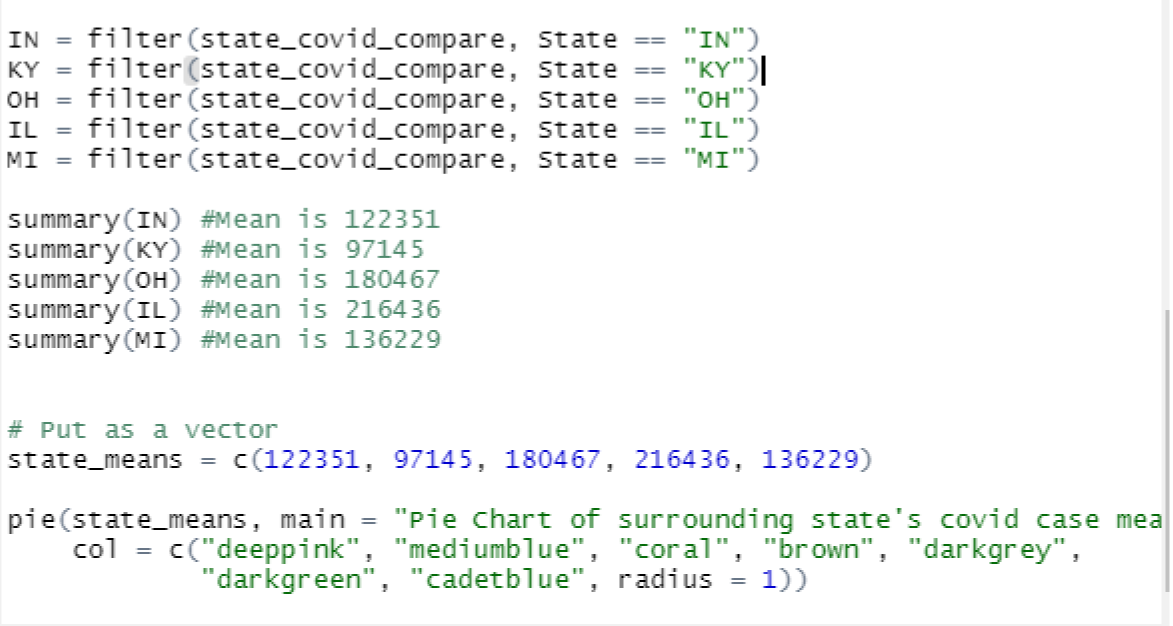


University	County	Year	Month	Monthly Cases
Bloomington	Monroe	2020	March	41
Bloomington	Monroe	2020	November	2402
Bloomington	Monroe	2020	December	2176
Bloomington	Monroe	2021	January	1526
Bloomington	Monroe	2021	June	185
Bloomington	Monroe	2021	September	1163
Bloomington	Monroe	2022	January	1,717
Bloomington	Monroe	2022	July	966
IUPUI	Marion	2020	March	1595
IUPUI	Marion	2020	November	20750
IUPUI	Marion	2020	December	23946
IUPUI	Marion	2021	January	13852
IUPUI	Marion	2021	June	1078
IUPUI	Marion	2021	September	12027
IUPUI	Marion	2022	January	48868
IUPUI	Marion	2022	July	7867
East	Wayne	2020	March	9
East	Wayne	2020	November	1851
East	Wayne	2020	December	1674
East	Wayne	2021	January	993
East	Wayne	2021	June	42
East	Wayne	2021	September	1770
East	Wayne	2022	January	4423
East	Wayne	2022	July	962
Kokomo	Howard	2020	March	39
Kokomo	Howard	2020	November	2341
Kokomo	Howard	2020	December	2702

This is the Excel dataset for counties and monthly cases

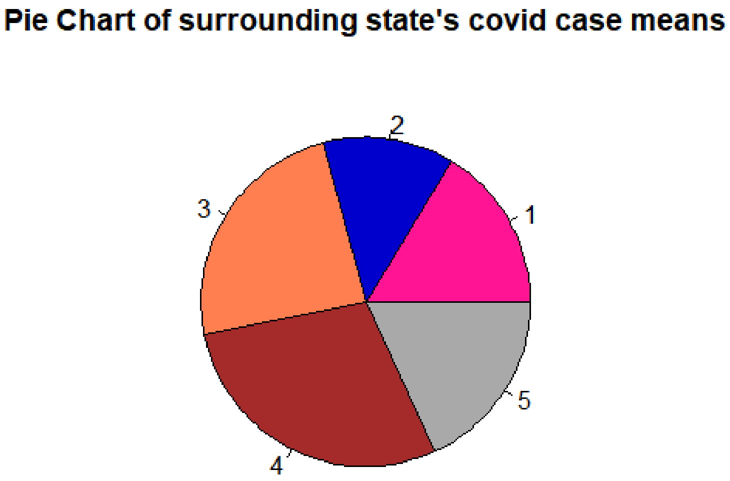
These datasets were created to be able to use simply in RStudio and be able to compare monthly case counts between the different counties and states.

Within RStudio the data was filtered by each county or state, such as the code visualizations below:



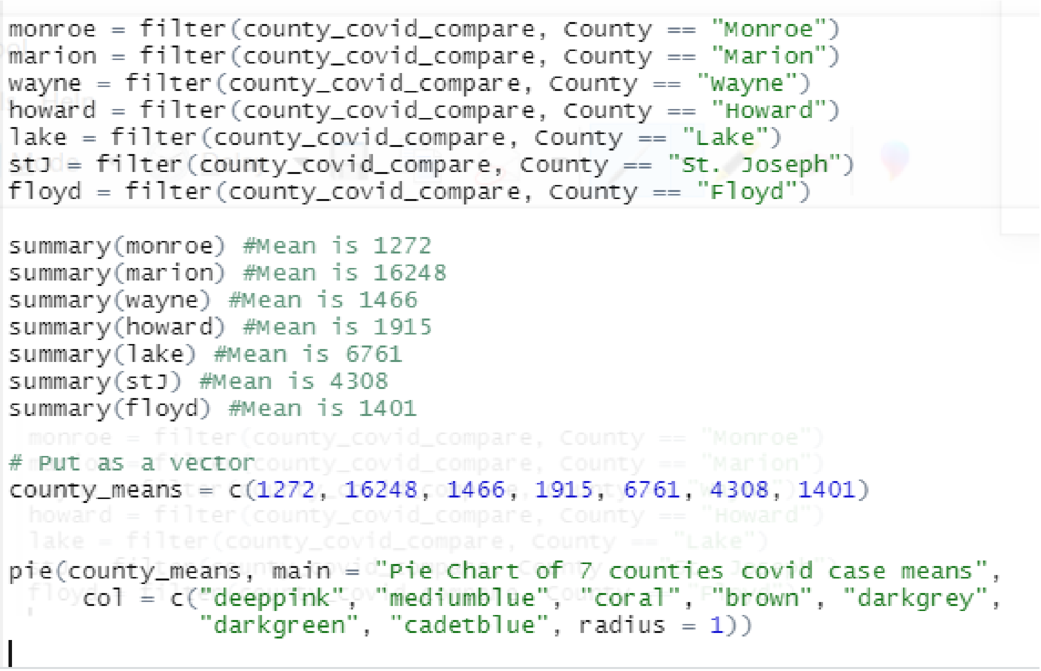
```
IN = filter(state_covid_compare, State == "IN")
KY = filter(state_covid_compare, State == "KY")
OH = filter(state_covid_compare, State == "OH")
IL = filter(state_covid_compare, State == "IL")
MI = filter(state_covid_compare, State == "MI")

summary(IN) #Mean is 122351
summary(KY) #Mean is 97145
summary(OH) #Mean is 180467
summary(IL) #Mean is 216436
summary(MI) #Mean is 136229


# Put as a vector
state_means = c(122351, 97145, 180467, 216436, 136229)

pie(state_means, main = "Pie Chart of surrounding state's covid case mea
    col = c("deeppink", "mediumblue", "coral", "brown", "darkgrey",
            "darkgreen", "cadetblue", radius = 1))
```

The graphic above shows code performed on the state dataset. We filtered each of the states from the main data, allowing a summary of each filter to be performed. It was decided the mean count would be best to use to showcase how each state differs in COVID cases. Each state's mean of COVID cases were then put into a vector to be used in a colored pie chart to showcase differences. Below is a visualization of the different states and their mean covid cases:



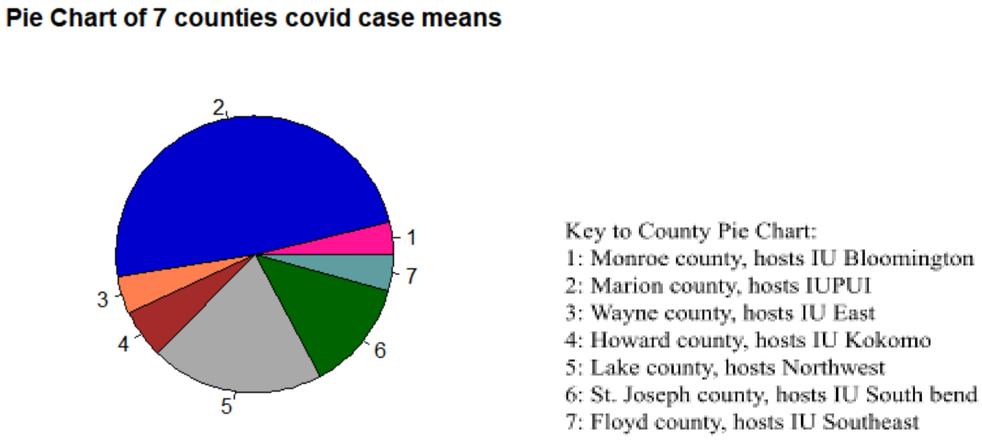
Key to State Pie Chart:
1: Indiana
2: Kentucky
3: Ohio
4: Illinois
5: Michigan

The above pie chart showcases the different states and their mean COVID cases. In this pie chart Illinois has the highest mean value of 216436, encompassing almost 25% of the chart. The next highest mean COVID cases go to Ohio and then Michigan, all state pies are nearly similar in size with some being slightly bigger. Drawing inferences from the dataset and pie chart, we can conclude that during November, December, and January (2020-2021) saw a rise in numbers while in the summer months (June, July, etc.) cases would lower. Other months/year where the numbers would rise as well such as new mutations being found, i.e. December 2021/January 2022 which saw the omicron variant being discovered and cases rose to the highest seen.



```
monroe = filter(county_covid_compare, County == "Monroe")
marion = filter(county_covid_compare, County == "Marion")
wayne = filter(county_covid_compare, County == "Wayne")
howard = filter(county_covid_compare, County == "Howard")
lake = filter(county_covid_compare, County == "Lake")
stJ = filter(county_covid_compare, County == "St. Joseph")
floyd = filter(county_covid_compare, County == "Floyd")

summary(monroe) #Mean is 1272
summary(marion) #Mean is 16248
summary(wayne) #Mean is 1466
summary(howard) #Mean is 1915
summary(lake) #Mean is 6761
summary(stJ) #Mean is 4308
summary(floyd) #Mean is 1401

# Put as a vector
county_means = c(1272, 16248, 1466, 1915, 6761, 4308, 1401)

pie(county_means, main = "Pie Chart of 7 counties covid case means",
     col = c("deeppink", "mediumblue", "coral", "brown", "darkgrey",
                "darkgreen", "cadetblue", radius = 1))
```

The graphic above shows code performed on the county dataset. We filtered each of the counties from the main data, allowing a summary of each filter to be performed. It was decided that the mean count would be best to use to showcase how each county differs in COVID cases. Each county's mean of COVID cases were then put into a vector to be used in a colored pie chart to showcase differences. Below is a visualization of the different counties and their mean covid cases:



The above pie chart showcases the different counties and their mean COVID cases. In this pie chart Marion county has the highest mean value of 16248, encompassing almost 50% of the chart. The next highest mean COVID cases go to Lake county and then St. Joseph county, all other counties have similar mean values. Drawing inferences from the dataset and information learned from the pie chart indicates Marion county (IUPUI) had the most COVID cases. Many counties had a lot of the same indications as the states dataset, where cases rose during/right after holidays and when new mutations are discovered.

Conclusion

The purpose of our research was to look at the trends of COVID-19 and what conclusions can be drawn from these trends. Utilizing the created datasets over states near Indiana and counties hosting the Indiana University campuses within R supported most of our hypotheses, including how during holidays and school starting/re-starting would show a rise in cases throughout all states (and thus the counties as well). However, one of our hypotheses was shown to be inconclusive, with bigger campuses in counties having more COVID-19 cases than smaller IU campuses. IU Bloomington has the most enrollments on average compared to other campuses but had one of the lowest mean counts and one of the smallest pie chunks on the pie chart. Research showed that Marion county, which hosts IUPUI, had the highest COVID cases compared to the other six counties with IU campuses in them. IUPUI has the second largest enrollment rate, after Bloomington, and had close to 50% of the pie chunk compared to other IU campuses. Most of the regional campuses (Kokomo, Southeast, East, etc.) have less than 10,000 student enrollment each year, however the counties that host IU Northwest and South Bend had the second and third highest COVID-19 mean counts compared to the high enrollment Bloomington. Some campuses are close to other states, such as IU East being close to Ohio state lines, IU Southeast being close to Kentucky state lines, IU Northwest being close to Illinois and Michigan state lines, and IU South Bend close to Michigan state lines. All states were relatively close to each in mean counts and were close in size on the state's pie chart.

This research into state traveling along with counties that host universities, specifically Indiana University campuses, could greatly benefit how schools decide to conduct learning, how bigger universities may impact spread of viruses compared to other smaller towns, etc. In 2020 with the onslaught of COVID-19, most states and universities went into a lockdown, online learning, and more precautions to help stop the spread. Since 2020 Indiana University has changed how learning is conducted, allowing courses that usually wouldn't be taught online to have the option of hybrid and online instruction.

Researchers could also look into the link between university towns and how those towns/counties could foster COVID-19 or any virus spread, especially towns near state border lines. The university age population is interesting to study as most present asymptomatic or have mild symptoms of COVID-19 which means this population can spread easier without knowing they have it and possibly mistake symptoms for something else. Undoubtedly, research into COVID-19 and its spread will help in any measure due to new mutations still being discovered and retransmission cases (second time or more COVID-19 suffers) rising.

Works Cited

Mangrum D., Niekamp P. (2022). JUE Insight: College student travel contributed to local COVID-19 spread. *Journal of Urban Economics*, 127(103311). Retrieved from: https://doi.org/10.1016/j.jue.2020.103311.

Gardner, L. (2020). Update January 31: Modeling the Spreading Risk of 2019-nCoV.

Johns Hopkins University CSSE. Retrieved from:

https://systems.jhu.edu/research/public-health/ncov-model-2/

Yilmazkuday, H. (2020). COVID-19 spread and inter-county travel: Daily evidence from the U.S. *Transportation Research Interdisciplinary Perspectives* 8(100244). Retrieved from: https://doi.org/10.1016/j.trip.2020.100244.

Hale, T., Atav, T., Hallas, L., Kira, B., Phillips, T., Petherick, A., Pott, A. Variation in US States' responses to COVID-19. *Blavatnik School of Government Working Paper.*

NCIRD (2022). United States COVID-19 Cases and Deaths by State Over Time. CDC, Public Domain U.S. Government. Retrieved from: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

Zlojutro, A., Rey, D. & Gardner, L. (2019). A decision-support framework to optimize border control for global outbreak mitigation. *Scientific Reports* 9(2216). Retrieved from: https://doi.org/10.1038/s41598-019-38665-w