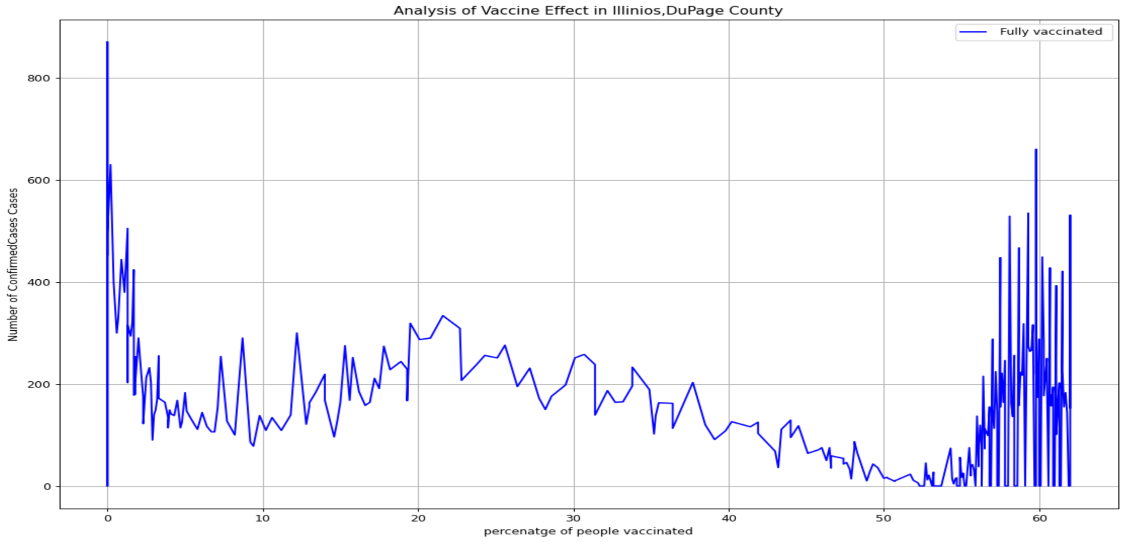
A7:Project Report

Name: Sindhu Madhadi
Project: " Analysis of Vaccines on Covid 19 Data for Illinois, DupageCounty"

1. Introduction

In this project, I wanted to analyze the impact of vaccination on the covid 19 cases in DuPage, Illinois County. I feel this is interesting and also important to know the change in the number of cases as people take vaccines, to see the effect of vaccine and how far it is protecting people. This problem infers to human-centered data science as it tells us how important the vaccines are from seeing the trends. To manage the spread and control of the pandemic, vaccination is important, knowing its effect in the number of cases helps us to take necessary actions against the spread of Covid.

By performing this analysis I tried to learn the trends in the Covid Cases for DuPage County. I aim to answer our questions and quantify the impact of Vaccines on Covid Cases in DuPage County by applying a combination of **Visualization graphs and predictive Testing** (Multiple linear regression).
The graph for the no of confirmed cases and Percentage of people who took the population is seen below.



Seeing the decrease in the case count I tried to dig deeper in my analysis and tried to find out trends for different age groups and also their correlation with county data.

2. Background/Related Work

During the pandemic, the cases started increasing nationwide. Nationwide lockdowns have been imposed to stop the spread. Also, vaccines have been released and given to people. Tracking changes in the number of cases with an increase in the percentage of people taking vaccines, give insights into the effectiveness of vaccines. The questions I tried to answer in this project are:

1. How significant is the data?
2. Does Vaccination decrease the count of confirmed covid cases?

Several articles have the same research to check the vaccine effects on covid. Here is one of the examples from [Reference 4].

Which has several methods and results for the vaccination effects in the US.

3. Methodology

The Analysis is carried out in three steps:

1. Data Acquisition - The data is downloaded from the below data source. The data set is large and has 32 columns and 1.09M rows. The columns in the data set are Date, FIPS, Country of Residence, State of Recipient, etc. the detailed list of 32 columns is found in the dataset link. Below is the table of the columns which I want to use in this scenario.

Column Names:

Date FIPS Series_Complete_pop_Pct(totoal population percentage) Recip_County(CountyName) Recip_State(state Name) Series_Complete_yes(Total population got vaccinated) Series_Complete_12Plus (population above !2+plus who got vaccinated) Series_Complete_12Pluspop_Pct (percentage of 12+ population who got vaccinated) Series_Complete_18Plus (18+ population who got vaccinated) Series_Complete_18PlusPop_pct (Percentage of 18+ population) Series_Complete_65Plus (65+ population who got vaccinated) Series_Complete_65PlusPop_Pct (percentage of 65+ population who got vaccinated) Metro_status (If it is a metro or non-metro)

2. Data Processing and Cleaning :

I have filtered data for DuPage County, Illinois, and merged the data with the existing Mask Mandate data on the date column. After Merging the two data sets ie Illinois DuPage Data and Covid Vaccines data based on date.

The two datasets are stored in the following files:

https://github.com/sindhumadhadi09/Data-512-Final-Project/tree/main/A6/Data_cleaned

3. Methodology:

Visualization Analysis:

Find the trends of Vaccination in COVID in Dupage Illinois

I tried to find the trends in data by different visualizations, ie by seeing the relation between vaccines taken by different age groups people and the number of cases in the county. The following are the visualization graphs, I plotted from the vaccinations data merged with the final data which we got from the Analysis of Mask Mandate and Illinois Data.

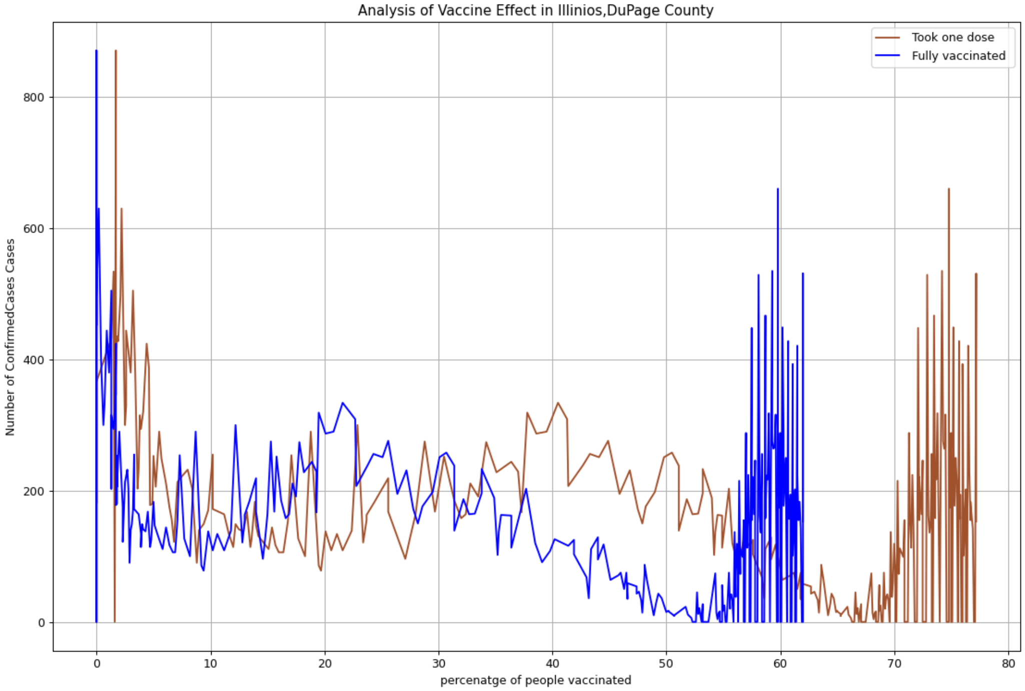
The graphs are plotted between:

No. of confirmed cases vs percentage of people vaccinated. (Graph Can be seen in Visualization result section)
No. of confirmed cases vs percentage of 12 plus people vaccinated. (Graph Can be seen in Visualization result section)

No. of confirmed cases vs percentage of 18 plus people vaccinated. (Graph Can be seen in Visualization result section)

No. of confirmed cases vs percentage of 65 plus people vaccinated.(Graph Can be seen in Visualization result section)

No. of confirmed cases vs percentage of people vaccinated vs people who have taken one dose.



Regression Analysis

I want to check the correlation of different age group of people vaccinated in our dataset, and their impact on the increase in Covid cases. For this, I choose Multiple Linear Regression as regression is best used to fund the correlation between different variables, as we have multiple variables to check I have used MLM. For this, I selected these 3 variables in the dataset for our analysis and checked whether they are statistically significant: Series_Complete_18PlusPop_pct Series_Complete_65PlusPop_Pct Series_Complete_12PlusPop_Pct

Call : lm(formula = DailyCases ~ Series_Complete_12PlusPop_Pct + Series_Complete_18PlusPop_Pct + Series_Complete_65PlusPop_Pct, data = data)

4. Findings:

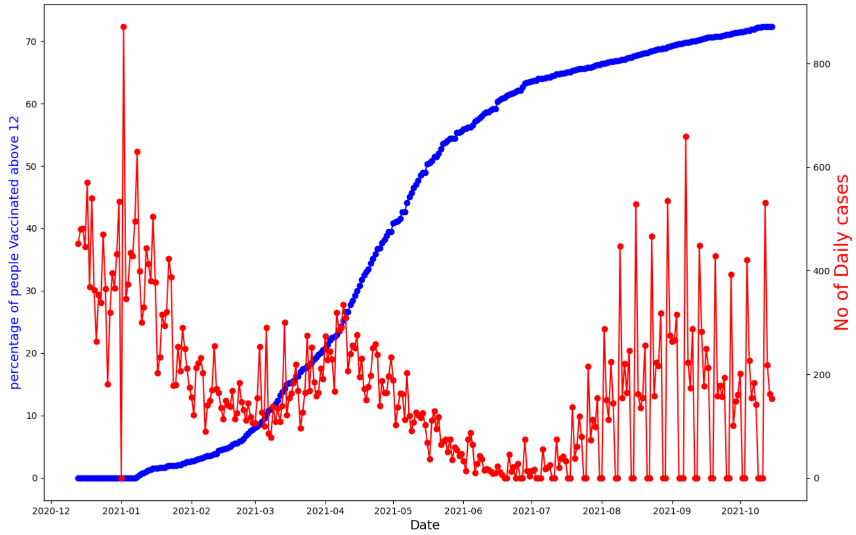
Visualization results

From the graphs, I see that there is a decrease in the Confirmed Covid cases with an increase in vaccination. It was not a continuous decrease but when we see on bigger picture on overall scenarios there is a decrease in confirmed cases. In the period 2021-08, we saw an increase in cases than 2021-07, but I think this can be my future work to explore more why in this period cases increased. I see the same results for all the three age groups of people:12plus, 18plus, 65plus.

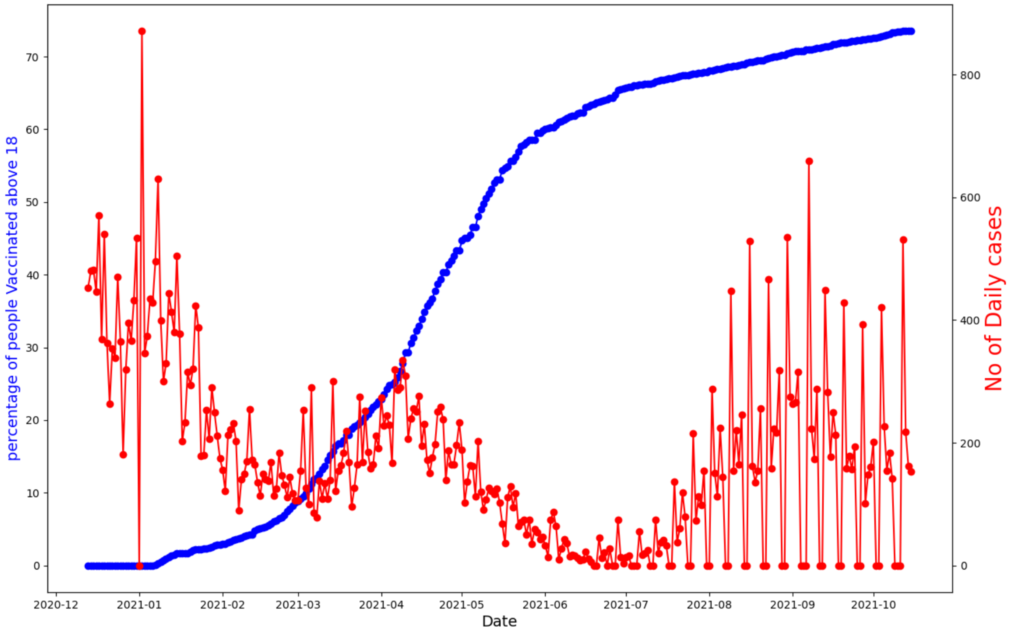
The graphs can be found in the link:

https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends_12Plus.png (12 Plus graph)

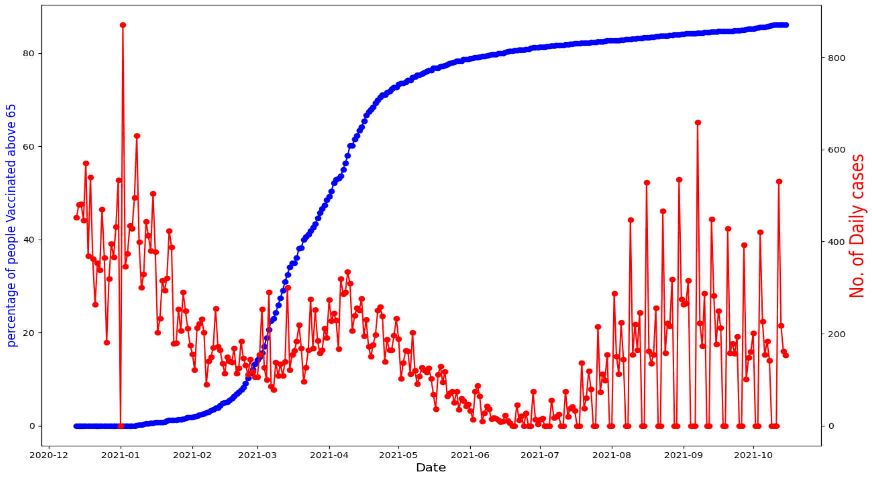
(12 plus along with date feature):



https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends_18Plus.png(18 plus along with date feature)



https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends_65Plus.png(65 plus along with date feature)



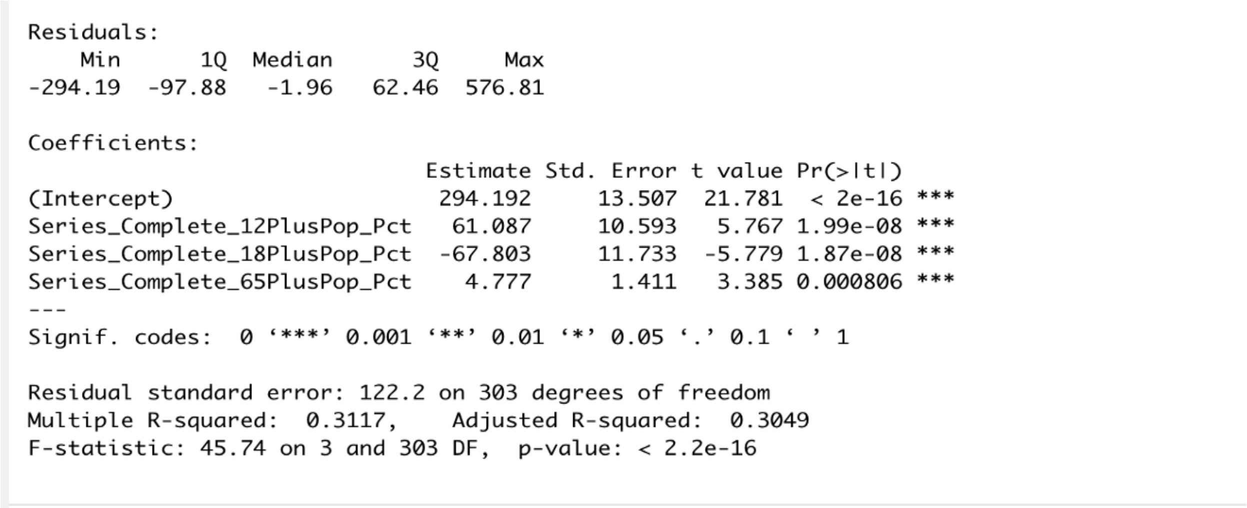
https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends.png(No. of confirmed cases vs percentage of people vaccinated.)

https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends_1.png(No. of confirmed cases vs percentage of people vaccinated vs people who has taken one dose.)

Regression Results

Based on p-values, if the p-value is less than alpha (0.05), then the variable is statistically significant. So, from the model, all the variables are statistically significant. Changes in these variables result in the change of Cases Count.

Based on the sign of the coefficient, we get whether it has a positive or negative correlation between the variables and the Cases Count.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-294.19  -97.88   -1.96   62.46  576.81 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    294.192     13.507  21.781  < 2e-16 ***
Series_Complete_12PlusPop_Pct   61.087     10.593   5.767 1.99e-08 ***
Series_Complete_18PlusPop_Pct  -67.803     11.733  -5.779 1.87e-08 ***
Series_Complete_65PlusPop_Pct    4.777      1.411   3.385 0.000806 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122.2 on 303 degrees of freedom
Multiple R-squared:  0.3117,	Adjusted R-squared:  0.3049 
F-statistic: 45.74 on 3 and 303 DF,  p-value: < 2.2e-16
```

5. Discussion/Implications:

It proved that seeing the trends that there is a decrease in cases with people taking vaccines.
However, there is a gap where I need to still further analyze why cases increased after 2021-08 compared to 2021-07.
There can be possibilities like there was no lockdown in that period or mostly during that period, or due to the cases where mak mandate was not imposed strictly.
Still, this needs to be figured out in the future work of the project.

For example, we can check the following graph:
https://github.com/sindhumadhadi09/Data-512-Final-Project/blob/main/A6/Output/vaccine_trends_65Plus.png(65 plus along with date feature)

6. Limitations:

A Few missing data points in the data are the current data set is from 12/13/2020 to 11/09/2021, our A4 common analysis is from February 1, 2020, through October 15, 2021.

So there is a gap in the period, I couldn't find a dataset exactly matching this period.

One more limitation is data in the given period has only metro cities. Also in the data, they could have given the details regarding the vaccine.

I tried multiple datasets but nowhere vaccine details are found, as that would help us, even more, to know which vaccines ins more effective and help to reduce the pandemic.

7. Conclusion:

From the graphs and regression, it is evident that with an increase in no of people taking vaccines there is a decrease in the case count in DuPage County

Also, from the regression, we found that the data is statistically significant.

8. References:

1. https://rpubs.com/rslbliss/r_mlm_ws
2. https://matplotlib.org/
3. https://stackoverflow.com/questions/5484922/secondary-axis-with-twinx-how-to-add-to-legend

4. .https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7709178/#:~:text=Vaccination%20markedly%20reduced%20adverse%20outcomes,respectively%2C%20across%20the%20same%20period.

9. Data Sources:

To perform the analysis I will use the following dataset.

1. **The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data - (A4 Analysis):**

   https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv

2. **MaskMandateDataset((in A4 Analysis):**

   https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

3. **The Vaccination Dataset - https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh**

The data set is large and has 32 columns and 1.09M rows. The columns in the data set are Date, FIPS, Country of Residence, State of Recipient, etc. the detailed list of 32 columns is found in the dataset link. Below is the table of the columns which I want to use in this scenario.

Column Names: Date FIPS Series_Complete_pop_Pct(totoal population percentage) Recip_County(CountyName) Recip_State(state Name) Series_Complete_yes(Total population got vaccinated) Series_Complete_12Plus (population above !2+plus who got vaccinated) Series_Complete_12Pluspop_Pct (percentage of 12+ population who got vaccinated) Series_Complete_18Plus (18+ population who got vaccinated) Series_Complete_18PlusPop_pct (Percentage of 18+ population) Series_Complete_65Plus (65+ population who got vaccinated)

Series_Complete_65PlusPop_Pct (percentage of 65+ population who got vaccinated) Metro_status (If it is a metro or non-metro)