**Covid Modeling Final Report**

**Team Members**

Raymond Adams (role: minder)

Fnu Shikha (role: grinder)

Ashutosh Godbole (role: grinder)

Pooja Kusuma Pillutla (role: finder)

**Project Description and Scope**

**Sponsors Understanding & Scope**

After the initial discussion with Adrian and Ines, it was understood that they are now looking at understanding the socio-economic / demographic factors of people and how it impacts the spread of Covid. If the spread is different for various factors, what is the explanation for this difference? They are also interested in building a Visualization that will aid in understanding the correlation between Socio-economic factors and number of cases, spread, immunity etc. Further analysis aims to determine that if at all there is indeed a correlation, can it be used to predict the spread of Covid in regions / areas with similar demographic structure. The second goal is to analyse the new COVID-19 variants and determine whether their survival is based on their own strength or outside factors.

**Teams’ Understanding & Scope**

Our teams’ awareness of the project goal was to find socio-economic factors that played a role in the spread of COVID-19. We also understood that once these factors were found, the correlation amongst each one was to be analyzed. As the team progressed throughout the project it was clear that not much public data containing desired and useful features were available for the latest variants of 2021. This caused the team to shift its research to data published in 2020. The team was able to find data on the earliest variants of 2020 that contained features such as: Average Case Density, Avg Infection Rate, Total No of Cases, Race, Ethnicity and Minority distribution etc. Once data was found our goal was to isolate the variables of choice and analyze their correlation to the spread of COVID-19. A conclusion was made for the work carried out by this team, but more analysis needs to be done to gain a better understanding of the characteristics of covid-19 variants and their differences amongst small communities that share similar socio-economic backgrounds.

**Data Availability and Documentation Repository**

Indiana State was selected for the analysis and hence all the data was pertaining to counties within the state of Indiana. Data was obtained from various sources and levels as given below:

**Covid Data:**

These datasets contain Covid Cases and the associated demographic and other data for those cases

* 1. **Covid-19 Case Surveillance file**:
     + *CDC\_COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography.csv*

This file is at individual case level / daily level and the most detailed/granular file. It is obtained from CDC website.

<https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/09172021.html>

* 1. **Covid\_report\_demographics:**
     + *hub\_mph\_covid\_report\_demo\_combined.xlsx*

This file is at County level. This file contains 4 tabs - Age, Race, Gender and Ethnicity. It contains Monthly data. This is fetched from hub\_mph site  
<https://hub.mph.in.gov/dataset/covid-19-case-data/resource/46b310b9-2f29-4a51-90dc-3886d9cf4ac1>

* 1. **Covid\_tracking\_Indiana\_combined\_Data:**
     + *Covid\_tracking\_Indiana\_Combined\_Data.xlsx*

This file is at State level (No county level data). Multiple Datasets have been merged to obtain this dataset. The site for getting this data is as below:  
<https://covidtracking.com/data/state/indiana>

* 1. **Covid\_Act\_Now:**
     + *counties.timeseries.csv*

This file is at County level and provides time series view.

The site for getting this data is as below:

<https://covidactnow.org/?s=25541467>

**Reference Data**

* 1. **Stats\_Indiana\_Edu**
     + *Indiana County level demographic data.xls*

This is a Reference dataset that provides demographic information for Indiana State Counties

This is sourced from below:

<https://www.stats.indiana.edu/topic/census.asp>

* 1. **Indiana Counties population 2020**
     + *Indiana Counties population 2020.xlsx*

This is a Reference dataset providing Area of the Counties in Indiana

* 1. **Stats\_America**
     + *Indiana County Metrics for development.xlsx*
     + *Indiana County Per Capita Income.xlsx*
     + *Indiana County\_Age\_Population.xlsx*
     + *Indiana County\_Employment.xlsx*

These are Reference datasets providing information at a County level on Age, Gender, Job Industry and Per Capita Income

<https://www.statsamerica.org/downloads/default.aspx>

All the datasets, documents, code file have been stored in Basecamp in the ‘Docs & Files’ section

<https://3.basecamp.com/3947469/buckets/21294307/vaults/3565110839>

**Approach and Work**

Spread of Covid is dependent on a number of factors such as Socio-economic, Demographic conditions, social measures (Lockdowns, social distancing, other restriction etc.), vaccinations as well as the Covid variant.

In order to study the impact of socio-economic and demographic factors, it is necessary to keep the other factors constant. At the beginning of 2020, when Covid started spreading, there were little to no social measures and there was only one known variant. Vaccinations were not yet produced / administered. Hence it easier to see if there is any correlation between spread of Covid and socio-economic factors.

We thus limited our analysis to the year 2020. In the first half of 2020, there were little to no social restrictions. However, second half saw various constraints being put on people’s movement by Governments of various States. Hence in this period, the spread of Covid would be impacted not only by social factors but also by the social measures

There were two levels of Visualizations

* **State**

Focussed on Indiana state level data

Only high-level demographic split of cases

No comparison done

* **County**
  1. County level analysis
  2. Demographic and Socio-economic data analysed
  3. Reference data such as Population and Area of counties added
  4. Derived attributes such as Population Density based on -> Area / Population
  5. Comparison of Counties
  6. Various Visualizations added to compare different attributes and whether they impacted the ‘Infection Rate’ or ‘Case Density’

The following Criterion were chosen to do a comparative analysis of the Counties

* Counties with High Case Density
* Counties with High Average Infection Rate
* Counties with High Population / Population Density

Since there are a number of counties with the State of Indiana, we chose top 10 or top 20 counties by various criteria and decided to analyse those.

**Results, Observations and Findings**

**Results:**

Page 1 - Shows the Counties in descending order for – Average Case Density (The number of cases per 100k population calculated using a 7-day rolling average), Avg Infection Rate, Avg Infection IC90 Rate, Total No of Cases

Page 2 – We take top 10 counties from page 1 based on Highest Case densities and plot:

* the timeline of their rise/fall in case density
* the timeline of their total no of cases

Page 3 – We take the top 10 counties with Highest case densities and plot

* + - distribution of race
    - distribution of housing occupied and vacant
    - distribution of Ethnicity (White Non-Hispanic, Minority)

Page 4 – We take the top 10 counties with Highest case densities and plot

* + - Distribution by Age Group
    - Distribution by Gender

Page 5 - We take the top 10 counties with Highest case densities and plot

* + - Distribution by Job Industry
    - Per Capita Income

Page 6 - We take the top 10 counties with Highest Infection Rate and plot

* + - 1. the timeline of their rise/fall in Infection Rate
      2. the timeline of their total no of cases

Page 7 - We take the top 10 counties with Highest Infection Rate and plot

* + - distribution of race
    - distribution of housing occupied and vacant
    - distribution of Ethnicity (White Non-Hispanic, Minority)

Page 8 – We take the top 10 counties with Highest Infection Rate and plot

* + - Scatter plot of ‘Full time work Index’ against ‘Literacy and Education Index’. The size of the bubble is proportional to the Case Density

Page 9 – We plot the top 20 counties by Population, Population Density and Total no of cases

Page 10 – We select top 20 counties bases on Total no of cases and plot

* + - Infection Rate vs Case density. The size of the bubble is proportional to the Population density

Page 11 - We select top 20 counties bases on Total no of cases and plot

* + - Infection Rate vs Case density. The size of the bubble is proportional to the Percentage of Minority Population

Page 12 - We select top 20 counties bases on Total no of cases and plot

* + - distribution of race
    - distribution of housing occupied and vacant
    - distribution of Ethnicity (White Non-Hispanic, Minority)

PDF File – *Covid - State level Visualisations.pdf*

This File uses the following datasets

* CDC\_COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography.csv
* hub\_mph\_covid\_report\_demo\_combined.xlsx

**Observations and Findings:**

* The First page lists the top 20 counties by each of the criteria. We see that the Average Infection Rate seems to be very close for each of these counties.
* Page 2 - We see that the rise in cases or timeline for Counties with Highest Case densities follow similar curve. ‘Cass’ County seems to follow a slightly different curve for both – Rise in Cases, Rise in Case Density
* Page 3 - The break of counties in terms of Race and Ethnicity/Minority also seems to be fairly even for the top 20. ‘Cass’ and ‘Clinton’ again stand out
* Page 4 - The Age distribution and Gender split are very similar across these 20 counties
* Page 5 - Dubois and Shelby have high Per Capita Income
* Page 6 - The rise in Infection Rates do not follow any particular pattern. All the counties considered have a different pattern
* Page 7 – Brown and Crawford seem to stand out slightly with the distribution of Housing data
* Page 8 – No particular pattern observed
* Page 9 – Counties with high population also seems to have high population density. For e.g., Marion, Lake, Hamilton, St Joseph, Allen, Vanderburgh figure in both High density and high population list
* Page 10 – County ‘Kosciusko’ seems to have low population density but high Case Density and High Infection Rate
* Page 11 – Where Minority Percentage of Population is more, the Infection Rate is quite high, but the case density is not high.
* Page 12 – Lake and Marion counties stand out with a different proportion of Race, Ethnicity and Minority distribution

**Conclusion**

* Counties with high Population also have high population density
* Vanderburgh Has high case density but low infection rate
* Marion, Lake counties should be studied thoroughly to get more information. These counties have population, high population density and also high infection rate
* The Case Density could be low and at the same time, Infection Rate could be high. For e.g., ‘Brown’ county
* ‘Elkhart’ county has high Infection Rate and High Case Density but low Population density. Hence this county needs to be studied for the reasons for high no of cases
* It is necessary to analyse the top 20 counties vs the bottom 20 counties for various criterion. This would help is checking how or whether some of the demographic factors have an impact
* County ‘Kosciusko’ to be studied as it has low population density, low population but still high Case density and infection Rate
* Counties like ‘Rush’ have high infection rate but low Case density and Population and hence they need to be studied thoroughly

**Recommendations for Project Continuation**

* We could not find Covid Case level data that had County information related to that particular case. It would be good to look for such data. This would be beneficial for analysis
* The project could not get detailed health related data (existing medical conditions) for the Covid cases. Recommendation is to search for this data and link it to the other datasets
* Try and link more demographic data (For e.g., Age Groups etc.) to Covid cases
* Continue the analysis with data for the year 2021
* Finding data pertaining to Covid Variants should be considered. This phase of the project worked upon building visualisations to study socio-economic impact on spread of Covid. However, combining that with the Covid Variants data would give a true picture