Analyzing 3 Chronic Diseases in the US: Diabetes, Asthma and Mental Health

Suman Shreyas
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x21169641@student.ncirl.ie

Rian Dwi Putra
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x22108637@student.ncirl.ie

Rohit Gopal Wadhwani
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x21194645@student.ncirl.ie

Abstract— this study will use ETL to combine data from an unstructured data format into a single repository that contains data that has been formatted correctly and qualified in advance of analysis. Access to additional processing and analysis is made simpler by this unified data repository. It likewise gives a solitary wellspring of truth, guaranteeing that all venture information is reliable and state-of-the-art.

Keywords—Python, ETL, Data Processing, Visualization

I. INTRODUCTION

This study focuses on 3 major chronic diseases which collected by the Center for Disease Control and Prevention (CDC) Division of Population Health, as follows.

1. Diabetes

A gathering of sicknesses that influence how the body utilizes glucose (glucose) is called diabetes mellitus. The muscles and other tissue cells get a lot of their energy from glucose. Additionally, it is the primary fuel source for the brain. The type of diabetes has a different primary cause. However, regardless of the type, diabetes can cause high blood sugar levels. A high blood sugar level can have serious effects on one's health.[3]

Figure 1. Measurement of blood glucose level

Both type 1 and type 2 diabetes are chronic forms of diabetes. Two types of diabetes can be reversed: pre-diabetes and gestational diabetes. Prediabetes is characterized by abnormally high blood sugar levels. However, the blood sugar levels are not sufficiently high to be considered diabetes. Furthermore, except if precautionary measures are taken, prediabetes can advance to diabetes. Gestational diabetes occurs during pregnancy. However, once the child is born, it might disappear.[5]

2. Asthma

When someone has asthma, their airways narrow, swell and may produce more mucus. Because of this, it may be difficult for them to breathe, which may result in coughing, wheezing when they exhale, and shortness of breath.[4]

For some, asthma can be a minor problem. For some, it can be a major problem that makes daily activities difficult and has the potential to trigger an asthma attack that could kill them.

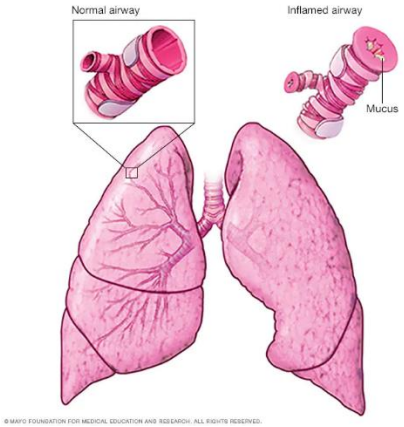

Figure 2. Asthma attack

Although there is no cure for asthma, its symptoms can be controlled. Asthma can change over time, so it's important to work with a doctor to monitor the signs and symptoms and adjust other treatments as needed.[6]

3. Mental Health

A wide range of mental health conditions is referred to as mental illnesses, which are also referred to as mental health disorders. These conditions affect a person's mood, thinking, and behavior. Sorrow, nervousness problems, schizophrenia, dietary issues, and habit-forming ways of behaving are instances of psychological instabilities.[7]

Numerous people experience mental health issues from time to time. However, a mental health issue becomes a mental illness when persistent signs and symptoms cause frequent stress and impair function.

Figure 3. Depression Symptoms

Psychological maladjustment can make somebody hopeless and cause issues in connections, school, work, or both. Symptoms can typically be managed with medication and psychotherapy (talk therapy).

The main objective of this study will interpret, are as follows:

1. Which Race / Ethnicity and Gender has the least immunity or is more affected by all these diseases?
2. Which State accounts for the most unvaccinated patients?
3. Which States have the most vaccinated Patients?
4. Which Vaccination has been sold the highest to patients and for which disease?

II. Related works

This research was carried out by CDC. The Centers for Disease Control and Prevention (CDC) is the most prominent public health service organization in the country that is driven by data and based on science. CDC has used science to help children stay healthy so they can grow and learn for more than 70 years; to assist families, businesses, and communities in maintaining their health and strength; and to safeguard the health of the public.

Asthma is currently a significant drain on healthcare resources in the countries studied thus far. Asthma has a significant impact on public health. Asthma affects approximately 18.5 million adults in the United States. In 2012, only 20% of high-risk individuals between the ages of 19 and 64 reported ever getting vaccinated against pneumococcal disease. [1].

Due in large part to the uncontrollable nature of asthma, its prevalence and severity will likely cause costs to rise. If disease control is improved, costs could be significantly reduced. The immediate costs of treating asthma control issues, such as hospitalization and trips to the trauma center, make up a significant portion of the total cost of the disease. When the disease is not completely controlled and becomes severe enough to affect daily life, indirect costs, such as time off from work or school and early retirement, are incurred. Quality of life assessments also demonstrate that asthma has a significant social and financial impact on patients and their families. Poor asthma control is significantly influenced by inadequate medication use and compliance. In asthma, noncompliance can increase healthcare costs, lead to additional complications, and in some cases, death. Worldwide guidelines recommending an expansion of the use of prophylactic treatment have been introduced to promote asthma across the board. Asthma-related hospitalizations will go down because of better asthma management, which may shift direct costs and, in turn, cut indirect costs. Working on useful interventions and engaging patient education programs can also help cut costs. This paper will examine the most recent study on asthma costs to determine how much of it is due to uncontrolled illness. Additionally, it will investigate areas where money could be saved.

In addition to asthma, diabetes and depression occur roughly twice as frequently together as would be predicted by chance alone. [2] Because their effects are exacerbated by one another, comorbid diabetes and depression pose a significant clinical challenge. This explanation fails to fully explain the connection between depression and diabetes, despite the possibility that the psychological burden of diabetes may make it worse. An inactive lifestyle, poor eating habits, inflammation, activation of the hypothalamic-pituitary-adrenal axis, disturbed sleep, and environmental and cultural risk factors are potential causes of both conditions. Despite the availability of effective screening tools, depression is frequently missed in diabetics. Depression in diabetics can be successfully treated with psychological interventions and antidepressants, but their effects on glycemic control vary. If people with diabetes and depression are to have the best possible medical and psychiatric outcomes, they need clear care pathways that involve a multidisciplinary team.

The public norms depend on proof of the advantages of DSME, which is a fundamental part of diabetes care. At the time of their diagnosis and as needed thereafter, Diabetes self-management education (DSME) should be provided to individuals in accordance with national standards. It helps diabetics maintain maximum level of self-care, which is necessary for maintaining a satisfactory quality of life, controlling complications, and maximizing metabolic control. It also helps to manage diabetes effectively when it is first diagnosed.

This study also measures days of mental illness. By combining recent physically and as a summary measure of population health, the mean number of mentally unhealthy days (days with impaired physical or mental health) over the previous 30 days is estimated. Social work is disrupted and associated with poor psychological well-being.

III. Methodology

1. Dataset Description

The CDI website serves as a gateway to additional data and information resources in addition to providing access to data on state-specific indicator variables.

The experiments presented in this study are based on every disease in the following datasets:

1. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Diabetes/f8ti-h92k
2. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj

3. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Mental-Health/ixrt-gnsg

Table 1. Disease Table Template

Column Name	Type
YearStart	Number
YearEnd	Number
LocationAbbr	Text
LocationDesc	Text
DataSource	Text
Topic	Text
Question	Text
Response	Text
DataValueUnit	Text
DataValueType	Text
DataValue	Text
DataValueAlt	Number
DataValueFootnoteSymbol	Text
DataValueFootnote	Text
LowConfidenceLimit	Number
HighConfidenceLimit	Number
StratificationCategory1	Text
Stratification1	Text
StratificationCategory2	Text
Stratification2	Text
StratificationCategory3	Text
Stratification3	Text
ResponseID	Text
LocationID	Text
TopicID	Text
QuestionID	Text
DataValueTypeID	Text
StratificationCategoryID1	Text
StratificationID1	Text
StratificationCategoryID2	Text
StratificationID2	Text
StratificationCategoryID3	Text
StratificationID3	Text

2. Detailed Description of Data Processing Algorithms

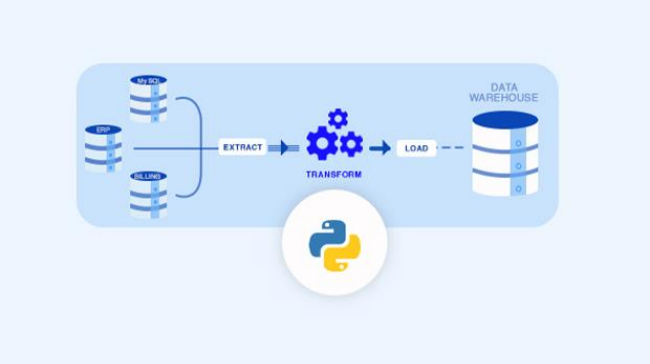

Figure 4. Data Processing Method

Data are extracted, transformed (cleaned, sanitized, or scrubbed), and loaded into an output data container in a process known as extract, transform, and load (ETL). The data can be compiled from many different sources and sent to many different places. ETL processing is typically carried out by hand, but administrators can also use software applications. The majority of ETL software do the entire process and run individually or collectively as a batch of jobs, either manually or on a recurring schedule.

Validity standards are enforced, data is extracted from the source system, and a well-designed ETL system ensures that the data meets the structural needs of the final output. Additionally, some ETL systems can deliver data in a ready-for-presentation format, facilitating the creation of applications by application developers and decision-making by end users. In this study, we use API as a data source for developing this project.

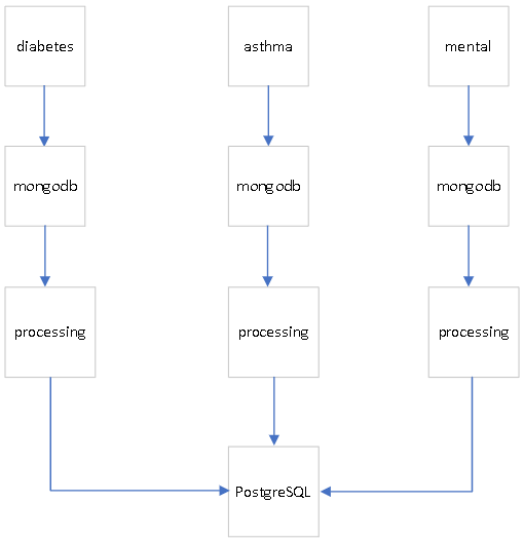

Figure 4. Migrating data from the source file to MongoDB and PostgreSQL

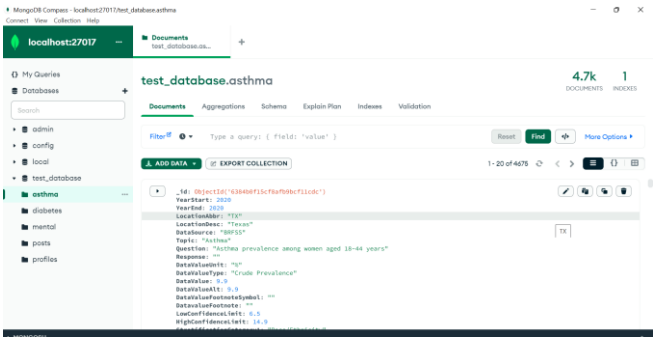

Figure 5. Migrated data in MongoDB Compass

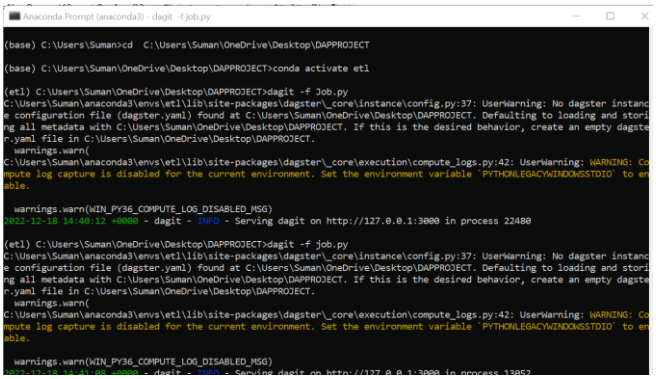

Figure 6. Executing the Job in Anaconda Prompt

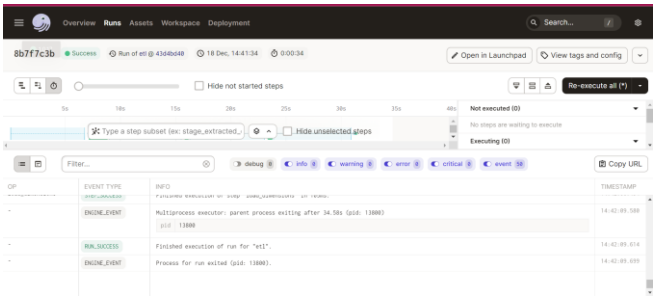

Figure 7. Execution of Job in Dragster web

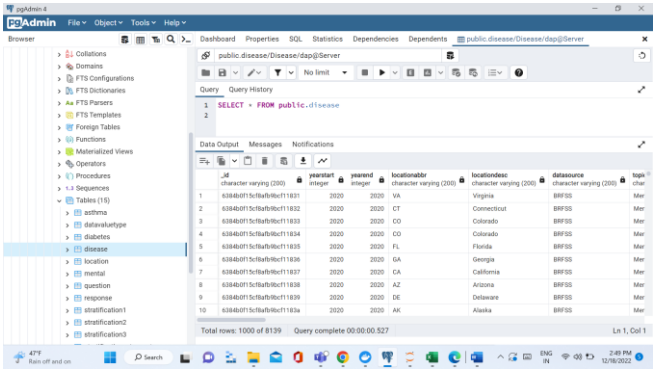

Figure 8. The migrated table in PgAdmin

Data migration is the process of selecting, preparing, extracting, and permanently transferring data from one computer storage system to another. To free up human resources from tedious tasks, data migration, which is a crucial part of this study along with implementation or consolidation, is typically carried out in a manner that is as clear as possible.

When migrating data, there are a few things to think about:

• Migration considers the possibility that the data carrier will become obsolete, but it does not take into account the possibility that the technologies that run the data will be abandoned completely, rendering migration useless.

• Consuming time: Each data object stored on a particular media must undergo migration, which is a continuous process that must be repeated whenever that media becomes obsolete.

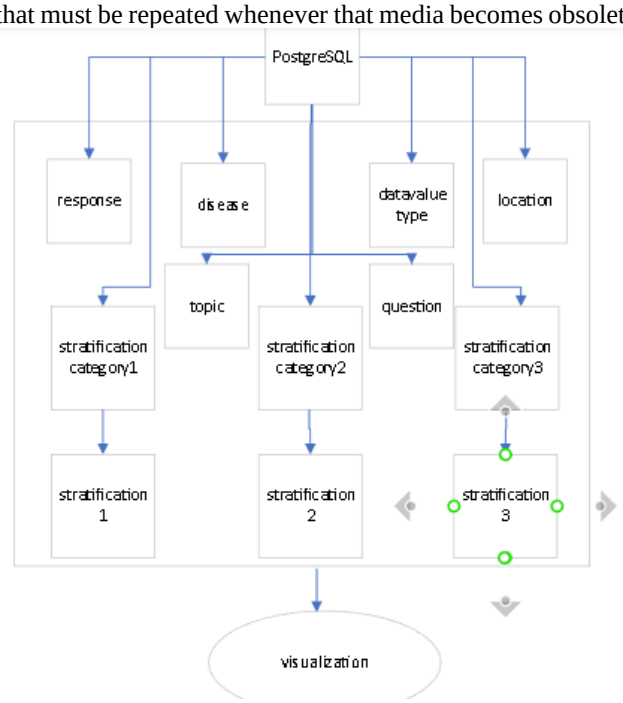

Figure 9. Distributed migrated data into a respective table in PostgreSQL

In the figure above, you can see the process of how the data is falling from PostgreSQL to Visualizations

```
SELECT distinct QuestionID, Question
INTO Question
FROM disease
order by QuestionID asc
```

Table 2. Question

Column Name	Type
QuestionID	Text
Question	Text

In this Table, we are showing the overall questions and the question ID by their Type

```
SELECT distinct TopicID, Topic
INTO Topic
FROM disease
order by TopicID asc
```

Table 3. Topic

Column Name	Type
TopicID	Text
Topic	Text

In this Table, we are showing the overall Topic ID and the Topics by their Type

```
SELECT distinct ResponseID, Response
INTO Response
FROM disease
order by ResponseID asc
```

Table 4. Response

Column Name	Type
ResponseID	Text
Response	Text

In this Table, we are showing the overall Response ID and the Response by their Type

```
SELECT distinct LocationID, LocationAbbr, LocationDesc
INTO Location
FROM disease
order by locationID asc
```

Table 5. Location

Column Name	Type
LocationID	Text
LocationAbbr	Text
LocationDesc	Text

In this Table, we are showing the Location ID, Location Abbr, and the LocationDesc by their Type

```
SELECT distinct DataValueTypeID, DataValueType
INTO DataValueType
FROM disease
order by DataValueTypeID asc
```

Table 6. DataValueType

Column Name	Type
DataValueTypeID	Text
DataValueType	Text

In this Table, we are showing the overall Data Value Type ID and the Data Value Type by their Type

```
SELECT distinct StratificationCategoryID1, StratificationCategory1
INTO StratificationCategory1
FROM disease
order by StratificationCategoryID1 asc
```

Table 7. StratificationCategory1

Column Name	Type
StratificationCategoryID1	Text
StratificationCategory1	Text

In this Table, we are showing the overall StratificationCategory ID1 and the StratificationCategory1 by their Type

```
SELECT distinct StratificationCategoryID2, StratificationCategory2
INTO StratificationCategory2
FROM disease
order by StratificationCategoryID2 asc
```

Table 8. StratificationCategory2

Column Name	Type
StratificationCategoryID2	Text
StratificationCategory2	Text

In this Table, we are showing the overall StratificationCategory ID2 and the StratificationCategory2 by their Type

```
SELECT distinct StratificationCategoryID3, StratificationCategory3
INTO StratificationCategory3
FROM disease
order by StratificationCategoryID3 asc
```

Table 9 . StratificationCategory3

Column Name	Type
StratificationCategoryID3	Text
StratificationCategory3	Text

In this Table, we are showing the overall StratificationCategory ID3 and the StratificationCategory3 by their Type

```
SELECT distinct StratificationID1, Stratification1
INTO Stratification1
FROM disease
order by StratificationID1 asc
```

Table 10. Stratification1

Column Name	Type
StratificationID1	Text
Stratification1	Text

In this Table, we are showing the overall StratificationID1 and the Stratification1 by their Type

```
SELECT distinct StratificationID2, Stratification2
INTO Stratification2
FROM disease
order by StratificationID2 asc
```

Table 11. Stratification2

Column Name	Type
StratificationID2	Text
Stratification2	Text

In this Table, we are showing the overall StratificationID2 and the Stratification2 by their Type

```
SELECT distinct StratificationID3, Stratification3
INTO Stratification3
FROM disease
order by StratificationID3 asc
```

Table 12. Stratification3

Column Name	Type
StratificationID3	Text
Stratification3	Text

In this Table, we are showing the overall StratificationID3 and the Stratification3 by their Type.

Now we will merge all the 3 files in one which will be done using SQL Query shown in the figure below

```
select * into disease from (
    select * from Asthma
    UNION all select * from Diabetes
    UNION all select * from Mental
)
```

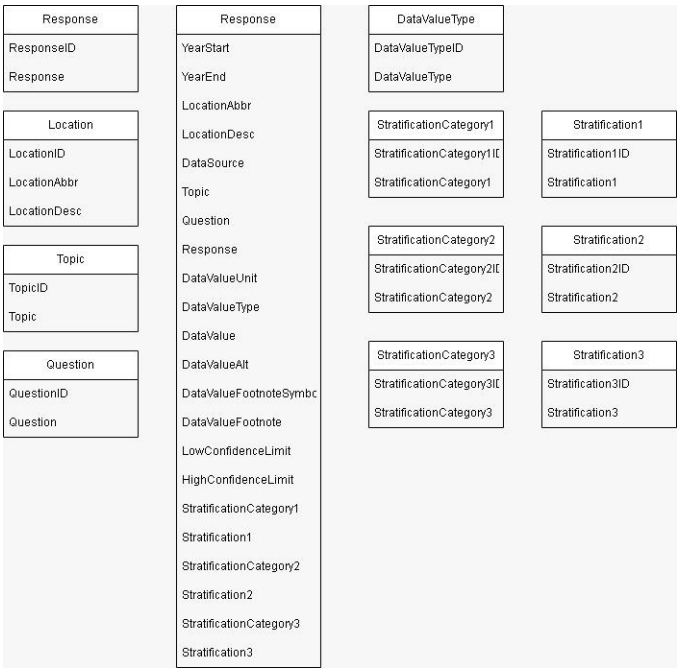

Figure 10. Database Schema

Here we can see the overall Database Schema of our project which includes all three diseases, where we can see the connection between the Disease table with all the other different tables.

The technology used in this study will focus on using Python, Docker Containers, and Dragster and for the database, MongoDB and PostgreSQL will be used to store the data. For analysis and visualization, this study will use python libraries and Jupyter notebook. The whole project source code is uploaded on GitHub for version control.

IV. Results and Conclusion

This study will analyze and visualize data using Python based on generated data.

A brief idea of every location and their averages, this is using the head function in Python to know our Data which we have shown the head and the tail of the data.

Table 13. Location and their average

	locationabbr	avg
0	CA	40.319380
1	NH	49.113208
2	OR	43.030973
3	US	47.946667
4	TX	41.260784

	locationabbr	avg
49	SC	42.925833
50	VA	47.217778
51	DE	47.701190
52	AZ	40.257047
53	KS	42.269841

Next would be Describing the data so we would get to know about the count of rows, mean, min, max, and all other descriptions.

Table 14. Dataset summary

	yearstart	yearend	datavalue	datavaluealt	lowconfidencelimit	highconfidencelimit
count	8139.0	8139.0	8139.000000	8139.000000	8139.000000	8139.000000
mean	2020.0	2020.0	40.531404	40.781128	33.956629	47.419941
std	0.0	0.0	27.118994	27.175552	24.430667	29.350988
min	2020.0	2020.0	2.000000	2.100000	1.400000	2.600000
25%	2020.0	2020.0	12.000000	12.300000	9.700000	15.500000
50%	2020.0	2020.0	41.300000	41.600000	31.900000	52.000000
75%	2020.0	2020.0	65.000000	65.300000	54.500000	74.100000
max	2020.0	2020.0	94.000000	94.000000	87.300000	97.700000

We have taken 2020-year data, in which we have a total of 8139 rows, including all 3 disease data, with a minimum value as low as 2.0 and a maximum of 94.

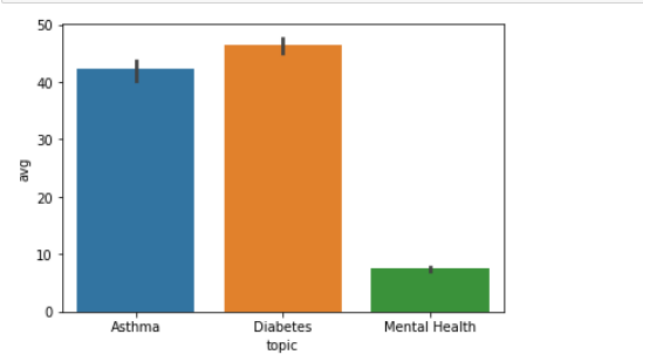

Figure 11. Average of Disease

Now we have taken the average of the disease individually and we can analyze and give insights that the mean of diagnosed diabetes that accounts for 46.91% which is higher than all other diseases, Asthma to be 42.67%, and Mental Health to be 7.63%.

Now in this dataset, we have 2 categories which include the percentage of people diagnosed without vaccination and Vaccinated people who were diagnosed.

In the below figure, we can see the states and the mean of the percentage of people who are diagnosed with the disease but are not vaccinated.

Table 15. Top 5 by location and where the topic is Diabetes

	locationdesc	topic	avg
0	Wisconsin	Diabetes	53.966667
1	North Dakota	Diabetes	52.767123
2	South Dakota	Diabetes	51.161765
3	Maine	Diabetes	50.816667
4	Delaware	Diabetes	48.382716

Table 16. Top 5 by location and where the topic is Asthma

	locationdesc	topic	avg
0	Wisconsin	Asthma	15.471429
1	Vermont	Asthma	14.800000
2	Rhode Island	Asthma	14.527778
3	Kentucky	Asthma	14.092857
4	West Virginia	Asthma	14.008333

By analyzing the table above, we can conclude that patients who are diagnosed with diabetes who are not vaccinated are the highest in Wisconsin with 53.96%, as compared to Asthma which has the least and accounts for the highest of 15.47%.

Now we will see the data for the patients who are vaccinated with the diseases. As we know there is no vaccination for mental health so only consider Asthma and Diabetes.

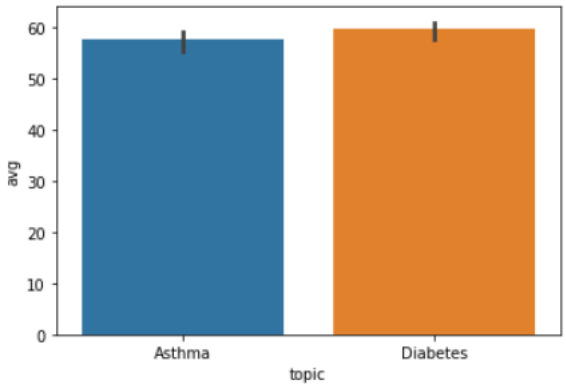

Figure 12. Average of disease with vaccinated patient

So, we can conclude that an average of more than 55% of patients are vaccinated for Asthma and Diabetes.

Now we would also analyze the Top 5 states which were vaccinated for both diseases.

Table 17. Top 5 location groups by topic

	locationdesc	topic	avg
0	Vermont	Diabetes	68.812500
1	New Hampshire	Diabetes	68.625000
2	South Dakota	Asthma	67.852941
3	Maine	Asthma	66.000000
4	Maryland	Asthma	65.952500

From Table 17, we can give insights that Vermont state has the highest percentage of patients who have taken vaccinated for Diabetes Disease and South Dakota has the highest percentage of patients who have taken vaccinated for Asthma Disease.

Table 18. Top 5 location

	locationabbr	avg
0	SD	54.475159
1	ME	53.627737
2	ND	52.504698
3	WI	52.456204
4	VT	50.081308

In the table, we can see the Top 5 locations which have the average of the highest percentage value for all the diseases. We can see that South Dakota has the highest mean of 54.47% then Maine with 53.62%.

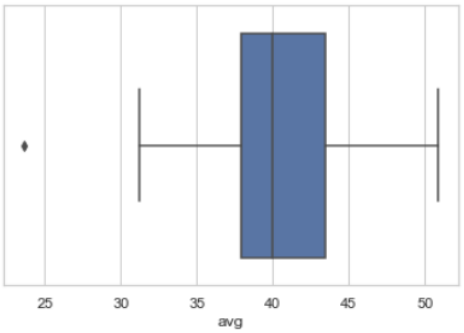

Figure 13. Mean of all diseases

In the previous Figure, we can see the mean of all the diseases and consider all the locations.

Table 19. List of all questions

	topic	datavalueunit	question	avg
0	Asthma	%	Influenza vaccination among noninstitutionaliz...	73.295897
1	Asthma	%	Influenza vaccination among noninstitutionaliz...	47.466403
2	Asthma	%	Asthma prevalence among women aged 18-44 years	13.294737
3	Asthma	%	Pneumococcal vaccination among noninstitutiona...	36.591411
4	Asthma	%	Pneumococcal vaccination among noninstitutiona...	83.329231
5	Asthma	%	Current asthma prevalence among adults aged >=...	10.391785
6	Diabetes	%	Visits to dentist or dental clinic among adult...	58.629496
7	Diabetes	%	Glycosylated hemoglobin measurement among adul...	67.369863
8	Diabetes	%	Prevalence of diagnosed diabetes among adults ...	10.774648
9	Diabetes	%	Influenza vaccination among noninstitutionaliz...	52.129921
10	Diabetes	%	Dilated eye examination among adults aged >= 1...	64.108108
11	Diabetes	%	Influenza vaccination among noninstitutionaliz...	70.157143
12	Diabetes	%	Diabetes prevalence among women aged 18-44 years	5.917293
13	Diabetes	%	Prevalence of depressive disorders among adult...	27.812992
14	Diabetes	%	Foot examination among adults aged >= 18 years...	69.993151
15	Diabetes	%	Pneumococcal vaccination among noninstitutiona...	40.618367
16	Diabetes	%	Adults with diagnosed diabetes aged >= 18 year...	50.432432
17	Diabetes	%	Pneumococcal vaccination among noninstitutiona...	75.622680
18	Mental Health	%	At least 14 recent mentally unhealthy days amo...	18.240529

In the figure above, we can see that during the survey they had asked these questions and we have taken the mean for each of the diseases, in which we can conclude that patients with Asthma had taken Influenza Vaccination has the percentage of people accounting a mean of 73.29%.

Now we would like to analyze which stratification category contributes to the diagnosis of all the diseases, so we

have taken the mean of the percentage of all the diseases with the stratification category.

Table 20. Average value per stratification category

	stratificationcategory1	avg
0	Overall	46.110771
1	Gender	49.212583
2	Race/Ethnicity	39.459191

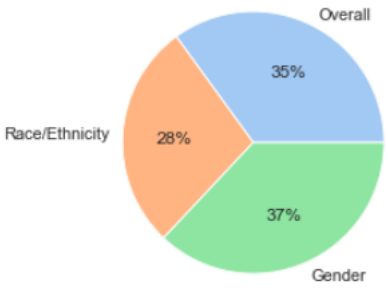

Figure 14. Average Diseases group by Stratification Category

By this, we can conclude that Gender has the highest percentage of patients who are diagnosed with all the diseases followed by an Overall 35%, and Race/Ethnicity. This is an overview of the percentage of patients diagnosed with all the diseases.

Now we will analyze more and try to find which Stratification has the highest percentage and all other stratification value has added up to a great percentage overall.

We would Start with Race / Ethnicity, so we have taken a pie chart to understand the percentage of all stratification which is contributing to Race / Ethnicity to 28%.

Table 21. Average value per Race/Ethnicity

	stratificationcategory1	stratification1	avg
0	Race/Ethnicity	Hispanic	27.914639
1	Race/Ethnicity	Black, non-Hispanic	37.794695
2	Race/Ethnicity	White, non-Hispanic	48.166167
3	Race/Ethnicity	Multiracial, non-Hispanic	19.956667
4	Race/Ethnicity	Other, non-Hispanic	29.202857

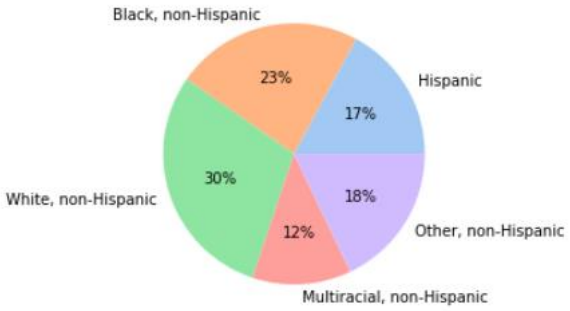

Figure 15. Average Diseases group by Race / Ethnicity

We can conclude that White, non- Hispanics have the highest percentage contributing 30%, and the mean of 48.16% which are diagnosed with all the diseases and followed by Black, non-Hispanics at 23% and all others.

Now we would check by Gender, As Gender only includes males and females, so in this, we will try to analyze which gender is affected more by all the diseases.

Table 22. Average value per gender

	stratificationcategory1	stratification1	avg
0	Gender	Female	48.504164
1	Gender	Male	43.148065

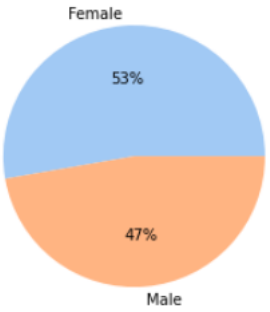

Figure 16. Average Diseases group by Gender

By looking at the table and Pie chart we can conclude that females are more prone to get infected with these diseases than men. As Female patients have contributed 53% and a mean value of 48.5%.

In the end, we can conclude that White Non- Hispanic Women are most affected by all these diseases than any other Race or Gender. For the states, we can conclude that the Wisconsin location has the highest number of unvaccinated patients, and the Vermont location has the highest percentage of vaccinated Patients.

V. FUTURE WORKS

This study concludes and interprets each diagram in the Result section based on the produced result.

If we have more time and opportunities to continue this project, this study is supposed to examine the Disease table in future works. This table can be used to draw some conclusions, such as.

- Distributed Disease in the United States by using Maps or geolocation
- We can create visualizations if we use another disease dataset that has more stratification, stratification category, data value unit, and/or data value type.
- We can create predictions or forecasting if we use a timeframe of more than 1 year period. For example, forecasting disease datasets based on a specific timeframe.
- Based on the CDC website, there are so many diseases we can use for research purposes. It means we can have more experiments using the dataset from this website.

VI. References

[1] P. J. Barnes, B. Jonsson, and J. B. Klim, “The costs of asthma,” European Respiratory Journal, vol. 9, no. 4, pp. 636–642, 1996.

[2] R. I. Holt, M. De Groot, and S. H. Golden, “Diabetes and depression,” Current diabetes reports, vol. 14, no. 6, pp. 1–9, 2014.

[3] “Diabetes,” Mayo Clinic, 07-Dec-2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444. [Accessed: 19-Dec-2022].

[4] “Asthma,” Mayo Clinic, 05-Mar-2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/asthma/symptoms-causes/syc-20369653. [Accessed: 19-Dec-2022].

[5] “Diabetes,” Who. int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 20-Dec-2022].

[6] “Asthma,” Who. int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma. [Accessed: 20-Dec-2022].

[7] “Mental disorders,” Who. int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. [Accessed: 20-Dec-2022].