**ViroShield**

**Big Data Processing Using Data Science Technology**

**Team MSG-Clusters**

**Maleeka Shahjahan (Student1281129)**

**Ahsan Siddiqui (Student1280746)**

**Abdul Samad (Student1285501)**

**Noor ul Ahad (Student1282756)**

# Abstract

Deaths often occur for unknown reasons, leading to various speculations. However, research highlights a significant rise in deaths caused by unnoticed virus outbreaks, posing a serious global issue. The World Health Organization (WHO) conducts crucial surveys to address such spikes in mortality and safeguard lives.

'Influenza' is a viral illness, showing symptoms like fever, headache, cough, and fatigue. Some symptoms last a week, but severe cases can be fatal. To tackle these challenges, a reliable application can be invaluable. Patients would be informed about their condition and initiate appropriate treatment promptly.

We did some research work for the data cleaning process. We analyzed the dataset so that we can recognize the metrics and key data. We tested different algorithms on the dataset to get the most accurate results. Big data analytics was a tough phase to cover where we did a lot of research work to get the best hands-on technology. We used Tableau, for data visualization although it was a bit tricky but we got the best out of it.

# Acknowledgment

We would like to express our gratitude to the Techwiz organizers for helping us by sending such assignments and organizing such competitions to boost our confidence and revision. We also want to express our gratitude to the Centre as a whole, our respected mentors and guides, and everyone who helped us with the project.

Additionally, our team members' dedication and relentless efforts have been truly commendable. They have invested significant time and expertise to ensure the successful development of the 'ViroShield' application. Their hard work has been pivotal in bringing together complex technologies, data analysis, and user-centric design to create a powerful tool that has the potential to save lives and make a positive impact on public health.

Table of Contents

[Abstract 2](#_Toc142917989)

[Acknowledgment 3](#_Toc142917990)

[Table of Figures 7](#_Toc142917991)

[What is Big Data? 8](#_Toc142917992)

[Introduction 9](#_Toc142917993)

[Problem Statement 9](#_Toc142917994)

[Functional Requirements 9](#_Toc142917995)

[Hadoop Cluster Data Loading 9](#_Toc142917996)

[Modeling Approaches 9](#_Toc142917997)

[Dataset Training, Validation, and Testing 9](#_Toc142917998)

[Leveraging Big Data Analytics 9](#_Toc142917999)

[Data Visualization 10](#_Toc142918000)

[Non-Functional Requirements 10](#_Toc142918001)

[User-Friendly Accessibility 10](#_Toc142918002)

[Seamless Integration 10](#_Toc142918003)

[Performance Efficiency 10](#_Toc142918004)

[Data Security 10](#_Toc142918005)

[Reliability: 10](#_Toc142918006)

[Research and Study 11](#_Toc142918007)

[Hadoop 11](#_Toc142918008)

[Hadoop Ecosystem 11](#_Toc142918009)

[HDFS (Hadoop Distributed File System) 11](#_Toc142918010)

[YARN (Yet Another Resource Negotiator) 12](#_Toc142918011)

[MapReduce 12](#_Toc142918012)

[Apache Spark 12](#_Toc142918013)

[Python 12](#_Toc142918014)

[NumPy: 12](#_Toc142918015)

[Pandas: 12](#_Toc142918016)

[Scikit-learn 13](#_Toc142918017)

[Flask 13](#_Toc142918018)

[SQLAlchemy 13](#_Toc142918019)

[Database 13](#_Toc142918020)

[MySQL 13](#_Toc142918021)

[Tableau 13](#_Toc142918022)

[Survey Information 14](#_Toc142918023)

[The Flow of the Project 15](#_Toc142918024)

[Downloading the Data 15](#_Toc142918025)

[Data Cleaning and Formatting 15](#_Toc142918026)

[Normalization 16](#_Toc142918027)

[Important Predictive Factors 17](#_Toc142918028)

[Design Specification 18](#_Toc142918029)

[Significance of the problem 18](#_Toc142918030)

[Utilization of Big Data 18](#_Toc142918031)

[Predictive Modeling 18](#_Toc142918032)

[Data Visualization and User Interaction 18](#_Toc142918033)

[Task Allocation 19](#_Toc142918034)

[Diagrams 20](#_Toc142918035)

[User Journey Map 20](#_Toc142918036)

[System Architecture 21](#_Toc142918037)

[Database Diagram 22](#_Toc142918038)

[Entity Relationship Diagram 22](#_Toc142918039)

[User Interface 23](#_Toc142918040)

[Sign Up Page 23](#_Toc142918041)

[Login Page 23](#_Toc142918042)

[Website Dashboard 24](#_Toc142918043)

[Previous Files 25](#_Toc142918044)

[Data Virtualization Dashboard on Tableau 26](#_Toc142918045)

[Main Dashboard 26](#_Toc142918046)

[Filtered Data by State 27](#_Toc142918047)

[Map View 27](#_Toc142918048)

[Comparison 28](#_Toc142918049)

[Values by Seasonal Percentage 28](#_Toc142918050)

[Dashboard Link: 29](#_Toc142918051)

[Conclusion 29](#_Toc142918052)

# Table of Figures

[Figure 1 Hadoop Architecture 10](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842126)

[Figure 2 Raw Data 13](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842127)

[Figure 3 Cleaned Data 13](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842128)

[Figure 4 User Journey Map 16](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842129)

[Figure 5 Hadoop Architecture 17](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842130)

[Figure 6 System Architecture 18](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842131)

[Figure 7 Hadoop Ecosystem 19](file:///C:\Users\Aazan\Downloads\Documentation%20(2).docx#_Toc142842132)

# What is Big Data?

**Big Data** refers to the vast and complex volume of data that exceeds the processing capabilities of traditional database systems. It encompasses both structured and unstructured data, often generated at high velocity and in immense quantities. Big Data is characterized by the three Vs: Volume, Velocity, and Variety. This concept has profound implications for various fields, including data science and projects like 'ViroShield.'

In the context of data science, Big Data represents the enormous datasets that require specialized tools, techniques, and frameworks for effective analysis. Data scientists work with Big Data to extract insights, patterns, and trends that can drive informed decision-making. Techniques like distributed computing, parallel processing, and advanced algorithms are used to process and analyze these datasets efficiently.

In the **'ViroShield'** project, Big Data plays a crucial role in handling the substantial amount of health-related data. With the proliferation of health information, including symptoms, virus spread, and patient records, traditional methods fall short. Big Data technologies enable the project to process and analyze these datasets comprehensively. This aids in predicting health issues, tracking virus patterns, and providing timely recommendations for individuals.

By leveraging Big Data techniques, **'ViroShield'** can handle the sheer volume and velocity of data generated during disease outbreaks. The project utilizes distributed computing frameworks like Apache Hadoop and Spark to process large datasets efficiently. Additionally, Big Data tools facilitate the creation of data visualizations, enhancing the project's ability to communicate insights to users and decision-makers.

In conclusion, Big Data is a paradigm that transforms how data science is conducted and how projects like 'ViroShield' tackle health-related challenges. It empowers data scientists and projects to unlock insights from immense datasets that were previously inaccessible, ultimately contributing to more informed decision-making and better public health outcomes.

# Introduction

## Problem Statement

Our top concern should be maintaining our health, which is one of the most crucial components of the modern world. A trustworthy application can be quite helpful in overcoming these obstacles. Patients would be made aware of their problem and swiftly begin the proper treatment. **'ViroShield'** is designed to assess users' overall health, particularly for potential virus infections. This aids users in understanding their condition and initiating treatment promptly, ultimately saving lives through accurate prescriptions. By utilizing the CDC dataset, we'll extract and train data to produce insightful results. Data visualization, like bar graphs, simplifies health status comprehension.

**'ViroShield'** aims to predict health issues and potential causes, tapping into the power of Big Data techniques. This technology efficiently processes and analyzes vast datasets, presenting health status on a user-friendly dashboard. The app generates predictions and showcases rising death cases or health concerns based on trained data. By leveraging advanced machine learning algorithms and distributed computing, **'ViroShield'** empowers individuals to proactively manage their health. This application stands as a testament to our commitment to harnessing cutting-edge technology to improve public health outcomes and save lives.

## Functional Requirements

### Hadoop Cluster Data Loading

* Load data into the Hadoop cluster, preparing it for subsequent analytics.

### Modeling Approaches

* Utilize data mining methods to predict values based on the dataset.

### Dataset Training, Validation, and Testing

* Train the acquired data by extracting, exploring, and analyzing relevant features.
* Perform thorough analysis to ensure accurate outcomes.

### Leveraging Big Data Analytics

* Capitalize on advancements in Big Data, benefiting from technologies like distributed computing and parallel processing.
* Leverage specialized frameworks like Apache Hadoop and Spark.
* Analyze data post-training to anticipate results showcased on a dashboard, enabling comparisons.

### Data Visualization

* Python offers a wide selection of packages and libraries designed for visualization purposes. We have also used Tableau for the visualization process.
* These tools empower users to create an array of visual representations, including plots, charts, graphs, and interactive visuals.
* The aim is to facilitate effective exploration and presentation of data.
* Dashboards play a key role by visually presenting essential data and metrics.
* This approach enables users to effectively monitor and gain a deep understanding of the Influenza situation.

## Non-Functional Requirements

Several non-functional requirements must be met by the application, including:

User-Friendly Accessibility

* The application should be user-friendly, ensuring effortless interaction with data.

### Seamless Integration

* The application should seamlessly integrate with diverse internal and external data sources to gather comprehensive Influenza-related data.
* Support for data integration from various applications, databases, and formats is essential.

### Performance Efficiency

* The application's performance should align with appropriate time constraints to produce accurate outputs.

### Data Security

* Data security is vital, necessitating encryption for protection.
* Unauthorized users must be prevented from downloading or utilizing data.

### Reliability:

* The application's reliability is crucial, ensuring consistent delivery of accurate outputs to end users.
* This includes predicted values and visualized data presented through bar graphs.

# Research and Study

Big data analysis is a trending topic in the information technology industry. We had to research and study a lot of content from the internet and books because we were not familiar with the topic. We studied different topics mentioned in SRS and were important to complete our tasks. Some main topics and technologies which we learn and used during this competition are mentioned below.

## Hadoop

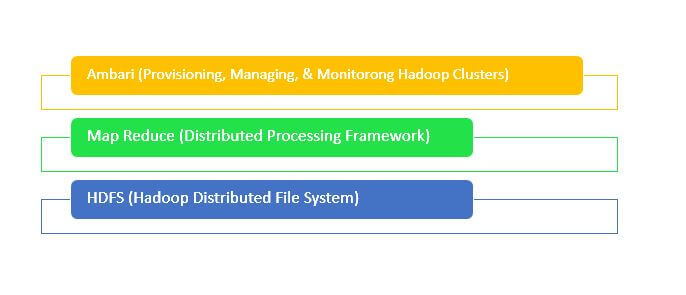
Hadoop is an open-source framework designed for processing and storing large datasets across distributed computing clusters. Hadoop enables fault tolerance, scalability, and efficient data processing, making it suitable for Big Data applications.

Figure 1 Hadoop Architecture

### Hadoop Ecosystem

The Hadoop ecosystem consists of two core components: HDFS (Hadoop Distributed File System) for data storage and MapReduce for parallel processing and analysis. Also, the ecosystem consists of many other components such as Casandra, Apache Spark, and so on. For the analysis of our data, we have used the components discussed below:

#### HDFS (Hadoop Distributed File System)

* Hadoop application data are stored using HDFS.
* The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications.
* Data from HDFS is consumed and processed using MapReduce applications.
* To construct a distributed file system that offers high-performance access to data across extremely scalable Hadoop clusters, HDFS uses a NameNode and DataNode architecture.

#### YARN (Yet Another Resource Negotiator)

* Apache Hadoop YARN is positioned between HDFS and the application-processing engines in a cluster design.
* YARN enables several data processing engines, including batch processing, stream processing, interactive processing, graph processing, and many more, to process and execute the data stored in HDFS (Hadoop Distributed File System).
* The efficiency of the system is increased with the use of YARN.

#### MapReduce

* A Hadoop cluster may scale massively with the help of the MapReduce programming paradigm over hundreds or thousands of machines.
* A programming approach called MapReduce is used for distributed, parallel processing of massive data collections.
* A programming approach called MapReduce is used for distributed, parallel processing of massive data collections.

#### Apache Spark

* Apache Spark is a freely available distributed processing platform employed for handling large-scale data tasks.
* Apache Spark serves as a data processing framework capable of efficiently executing tasks on extremely vast datasets.
* It employs memory caching and optimized execution of queries to swiftly process analytical queries on data of varying sizes.

## Python

Python is a versatile, high-level programming language known for its readability and simplicity. Its rich ecosystem and ease of use make Python a popular choice for data analysis, scripting, and web development. Python offers a wide range of libraries and modules and we have used some of them, which are as follows:

### NumPy:

NumPy library for handling numerical data efficiently and numerical computation.

### Pandas:

It's used to create a two-dimensional labeled data structure called a DataFrame, which is one of the core data structures in Pandas. It’s also used for data manipulation.

### Scikit-learn

This model is used for machine learning. It splits the dataset into training and testing subsets, a crucial step in building and evaluating machine learning models.

## Flask

Python is used to create the Flask web-based application framework. It was created by Armin Ronacher, who served as the team leader of Poocco, an international group of Python enthusiasts. The Werkzeg WSGI toolkit and the Jinja2 template engine are the foundation for Flask. They're both Pocco projects.

### SQLAlchemy

* SQLAlchemy is an open-source SQL toolkit and object-relational mapper for the Python programming language.

## Database

The important data of the application will be saved in the database so users can go through data uploads. Databases are the essential data repository for all software applications. For example, whenever someone conducts a web search, logs in to an account, or completes a transaction, a database system is storing the information so it can be accessed in the future.

### MySQL

* MySQL is a relational database management system.

## Tableau

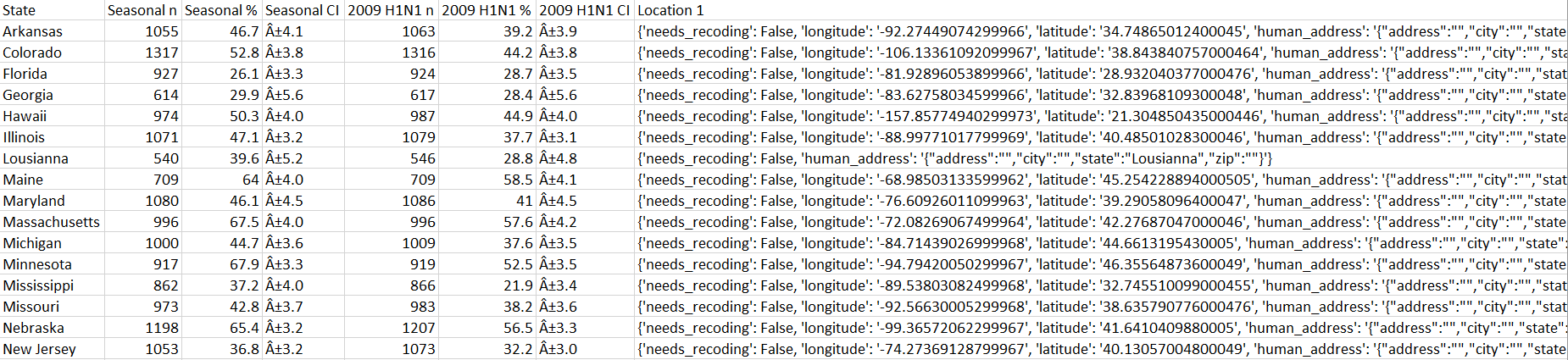
* Tableau is a leading data visualization tool that helps users create interactive and insightful visual representations of data.
* It offers a drag-and-drop interface to design various types of charts, graphs, maps, and dashboards without requiring complex coding.
* Tableau's interactive features enable users to explore data, identify patterns, and gain insights from complex datasets.
* We have used maps, bar charts, and graphs for the detailed visualization of the prediction of our cases.

# Survey Information

The sample data set which was given in the requirements was from the **CENTERS FOR DISEASE CONTROL AND PREVENTION** titled, **State Specific Influenza Vaccination Coverage** which contains the state-wise data coverage to give the different attributes like state's name, zip code, 2009 C1H1, 2009 confidence interval, and so on.

The **Socrata API** and the **Kaggle API** are used to maintain this dataset. Socrata has helped innumerable organizations host their open data and has played a crucial role in making more data available to the public.

Data set link: <https://www.kaggle.com/datasets/cdc/state-specific-influenza-vaccination-coverage?select=state-specific-influenza-vaccination-coverage-among-women-with-live-birth-prams-2009-10-influenza-season.csv>



# The Flow of the Project

We followed the given flow and systematically assessed performance step by step as outlined below.

## Downloading the Data

* We initiated the project by addressing raw data in its unprocessed form.

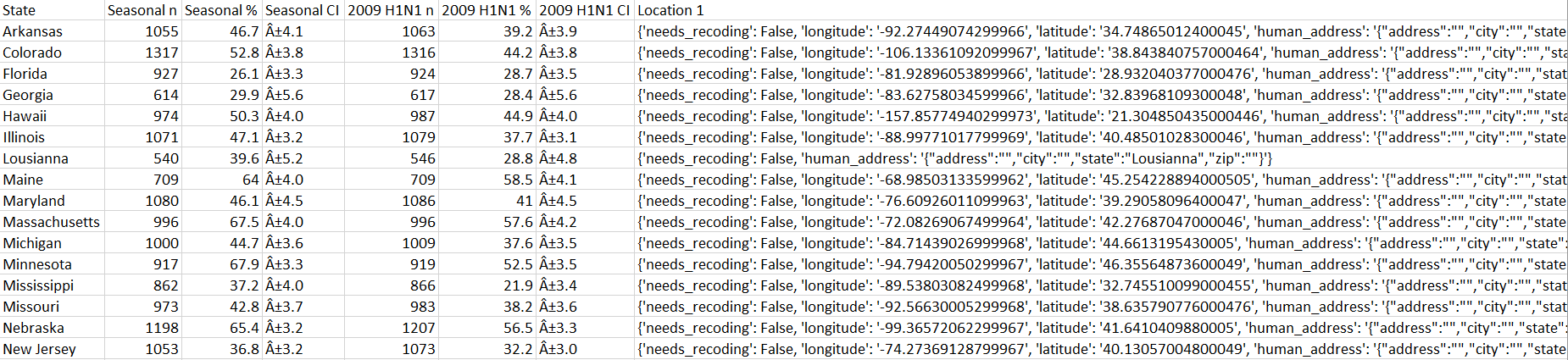


Figure 2 Raw Data

## Data Cleaning and Formatting

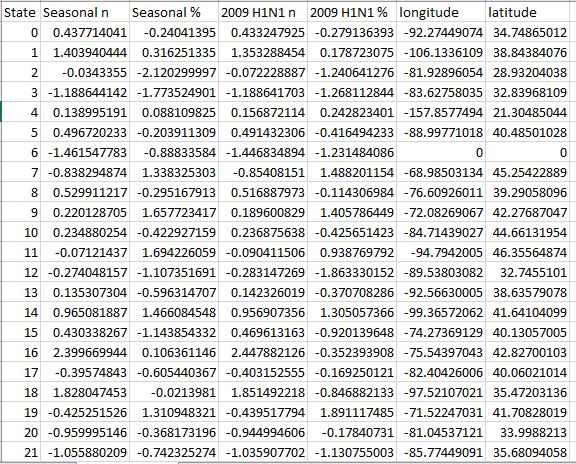
* The data cleaning phase aimed to rectify inconsistencies and outliers. We used a combination of Python’s libraries Pandas and Scikit-learn. It was decided that some fields are not important for data processing so we filter out those fields.

Figure 3 Cleaned Data

* The discrepancy you're observing in the "Seasonal n" column's value transformation from "1055" to "0.437714040606874" is due to the normalization step that was performed in the data you provided.
* The human location data has been extracted from the values of longitudes and latitudes.
* Following data preprocessing, we seamlessly transitioned to the critical phase of data modeling and training. Our approach harnessed the power of big data analytics, using advanced algorithms and distributed computing to uncover insights.
* We selected algorithms known for handling large datasets efficiently, capitalizing on frameworks like Apache Hadoop and Spark for parallel processing.
* Having developed and fine-tuned our models during the training phase, the prediction phase is where we leverage these models to make informed predictions on new or unseen data. This phase capitalizes on the insights gained from the trained models to forecast outcomes, trends, or classifications.
* The user will upload the file from the web interface created. The file then go to HDFS system, it process it and then returns preprocess data.
* Then Python algorithm will then analyze the data and predict the results.
* Once predictions have been generated, the analyzing phase comes into play. This phase involves the thorough examination and interpretation of the predictions and their implications.
* Analyzing the predicted outcomes is critical to extracting meaningful insights and actionable intelligence. Predictions alone don't provide the complete picture. The analyzing phase involves understanding the context and significance of the predicted outcomes.

## Normalization

Normalization is a common preprocessing technique in data analysis, especially when you're dealing with numerical attributes that have different scales. The purpose of normalization is to scale all numerical attributes to a common scale, usually between 0 and 1 or with a mean of 0 and a standard deviation of 1. This helps algorithms that rely on distances or gradients, such as machine learning models, to perform better.

# Important Predictive Factors

Prediction is the core requirement of our project. The data set cleaned by Python was predicted using Scikit-learn. The data prediction has entirely consisted of the factors selected for prediction. The column "State" is used as a feature column and the predicted column was "2009 N1H1 n". The average is used for "2009 N1H1 n".

When the prediction is completed, we will change the values from the predicted data to the original data to filter out the data by "State".

# Design Specification

The 'ViroShield' project aims to develop an application that utilizes Big Data techniques, predictive modeling, and data visualization to address the challenge of influenza-related deaths. This research analysis explores the significance of the project, its potential impact, and the technologies involved.

## Significance of the problem

Influenza remains a global health concern, causing significant morbidity and mortality each year. Unknown and undiagnosed cases of influenza can lead to delayed treatment, contributing to the rise in deaths. The proposed 'ViroShield' application addresses this issue by leveraging Big Data analytics to predict influenza outcomes, enabling early intervention and treatment.

## Utilization of Big Data

Big Data technologies offer the capability to process and analyze vast amounts of data efficiently. By utilizing frameworks like Apache Hadoop and Spark, the application can handle large and diverse datasets, enabling accurate predictions and data-driven insights.

## Predictive Modeling

The core of the application involves data mining and machine learning techniques for predictive modeling. By training models on historical and real-time influenza data, the application can forecast health outcomes, including identifying potential fatal cases. This proactive approach empowers individuals and healthcare providers to take timely actions, reducing mortality rates.

## Data Visualization and User Interaction

Data visualization is a powerful tool for conveying complex information in an understandable format. The integration of Tableau Desktop allows the application to create interactive dashboards with various visualizations such as bar graphs and charts. These visual representations enhance user understanding and enable effective exploration of health data.

# Task Allocation

|  |  |  |
| --- | --- | --- |
| Task | Member | Status |
| Research and study about Hadoop, Apache Spark, Big Data, and other important concepts | Noor, Ahsan, Samad, Maleeka | Completed |
| Configure Hadoop | Noor, Ahsan, Maleeka | Completed |
| Apache Spark configuration | Noor, Ahsan | Completed |
| Tableau download and dashboard designing | Samad | Completed |
| Map work on Tableau | Samad | Completed |
| User interface designing | Ahsan | Completed |
| Backend development on Flask | Ahsan, Noor | Completed |
| Hadoop HDFS file system, prediction with Python libraries like Scikit-learn, and others | Noor | Completed |
| Verify the predictions, and validate algorithms | Noor | Completed |
| Documentation and research | Maleeka | Completed |
| Testing | Maleeka | Completed |

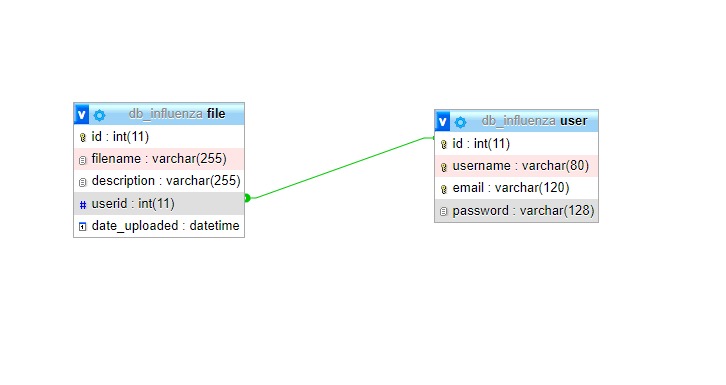
# Diagrams

## User Journey Map

## System Architecture

This is the system architecture of the project. First, the data was downloaded and then applying preprocessing data techniques then the data will be analyzed and the output will be shown to the user through visualization.

## Database Diagram



As per the requirements, the user can upload the files and see their data through virtualization but the data needs to be stored so users can see the previous uploads. For this purpose, the MySQL database has been used in this project.

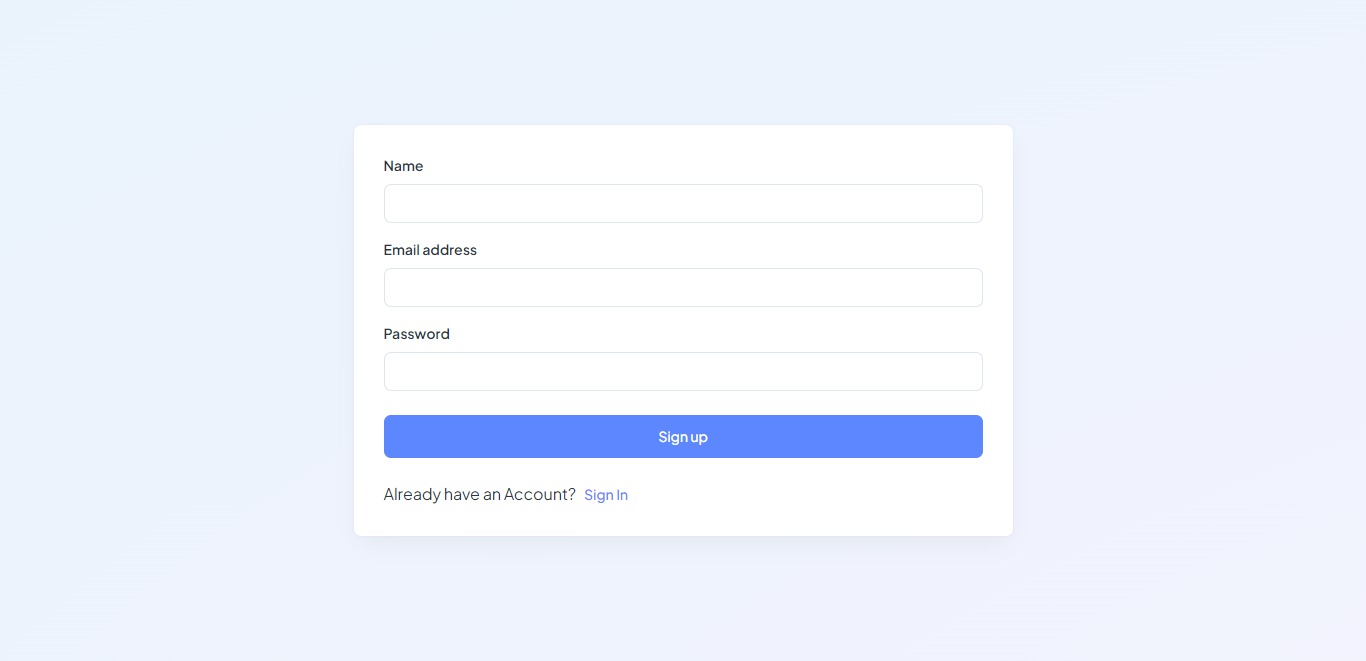
## Entity Relationship Diagram



An entity–relationship model describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types and specifies relationships that can exist between entities.

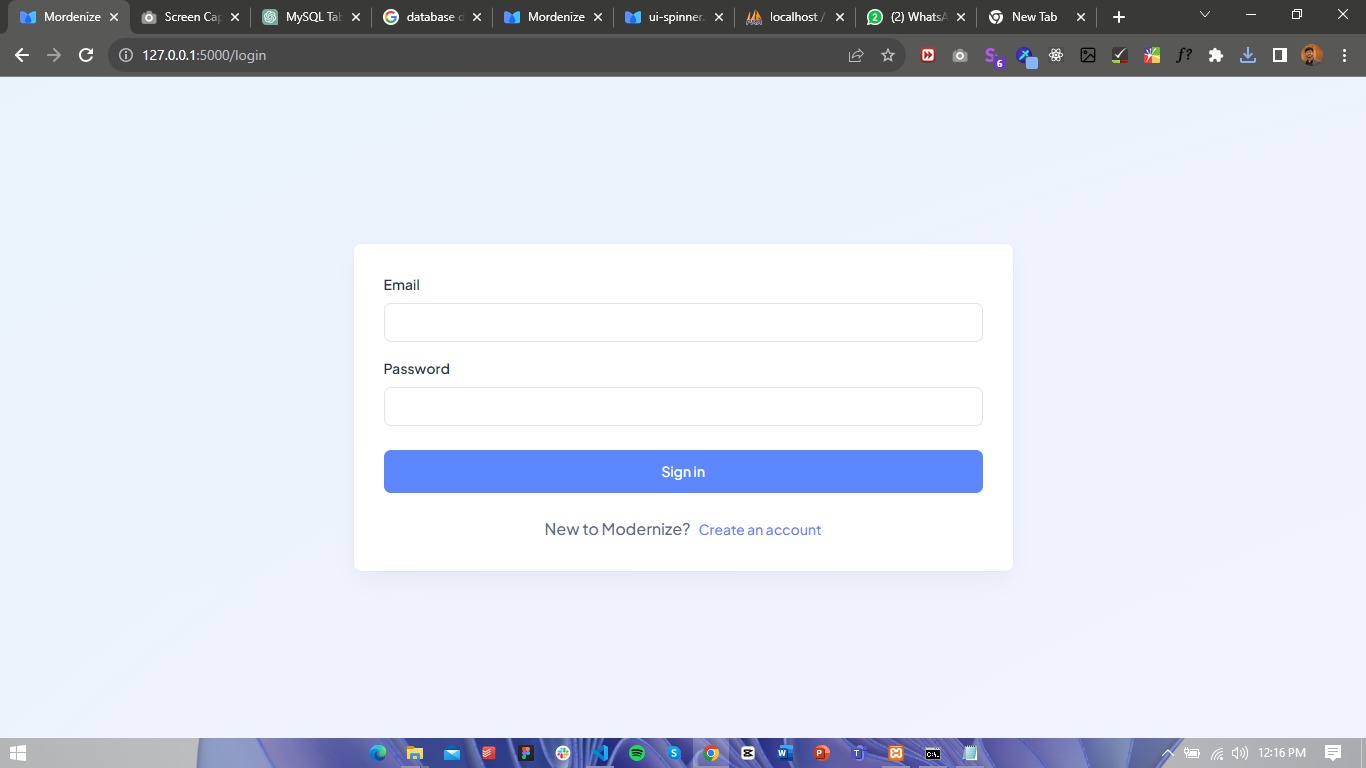
User Interface

## Sign Up Page



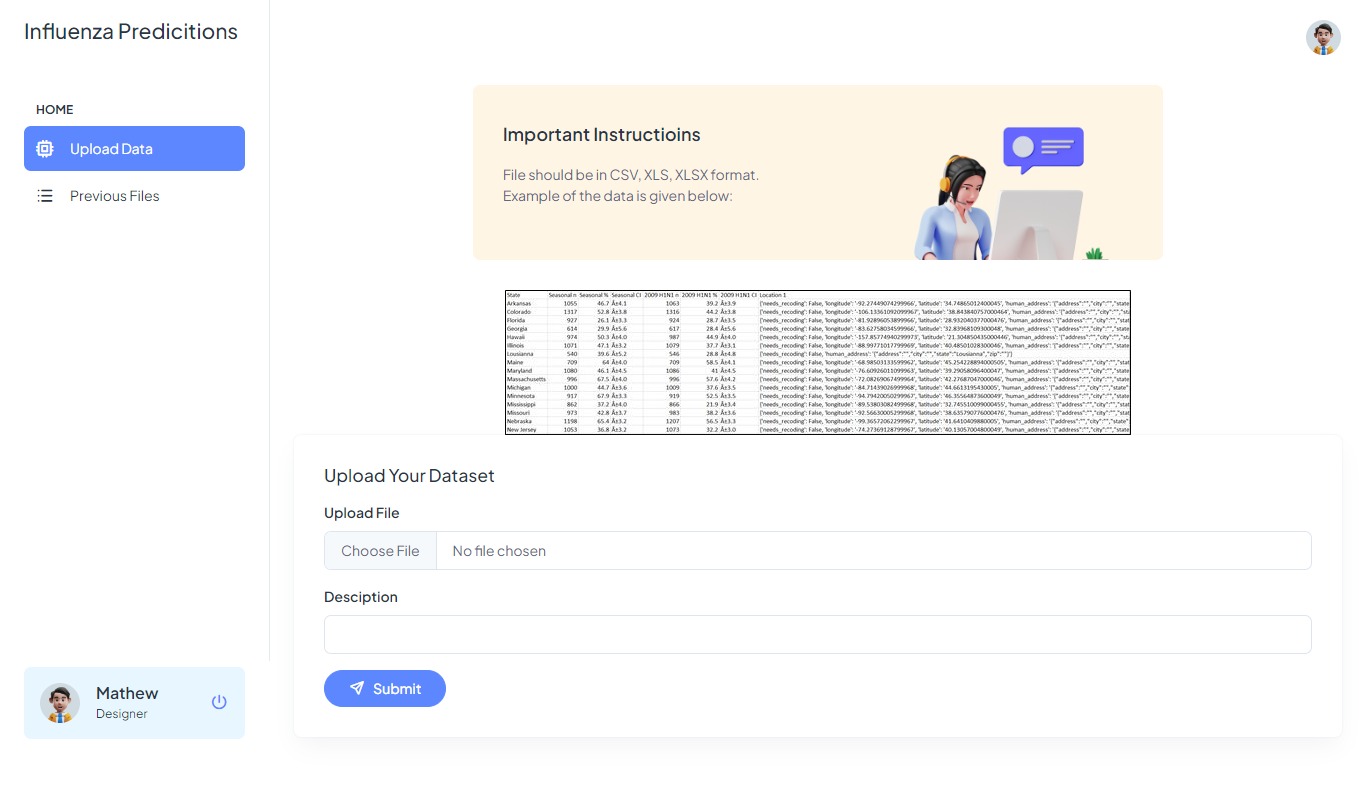
From this page you can sign up as a user.

### Login Page



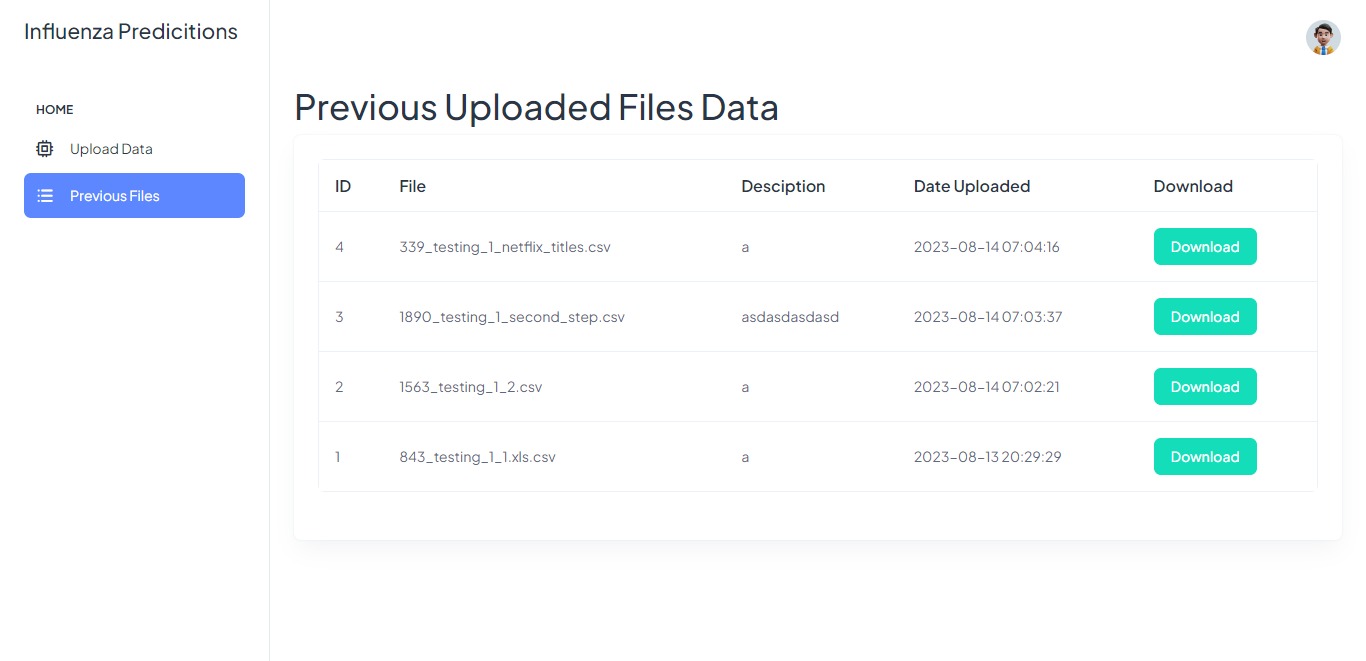
From this page you can login as a user and visit dashboard to upload the files and see previous uploaded data.

### Website Dashboard



From this page, you can upload your file for predictions. The instructions are mentioned above so you will know the file format.

### Previous Files

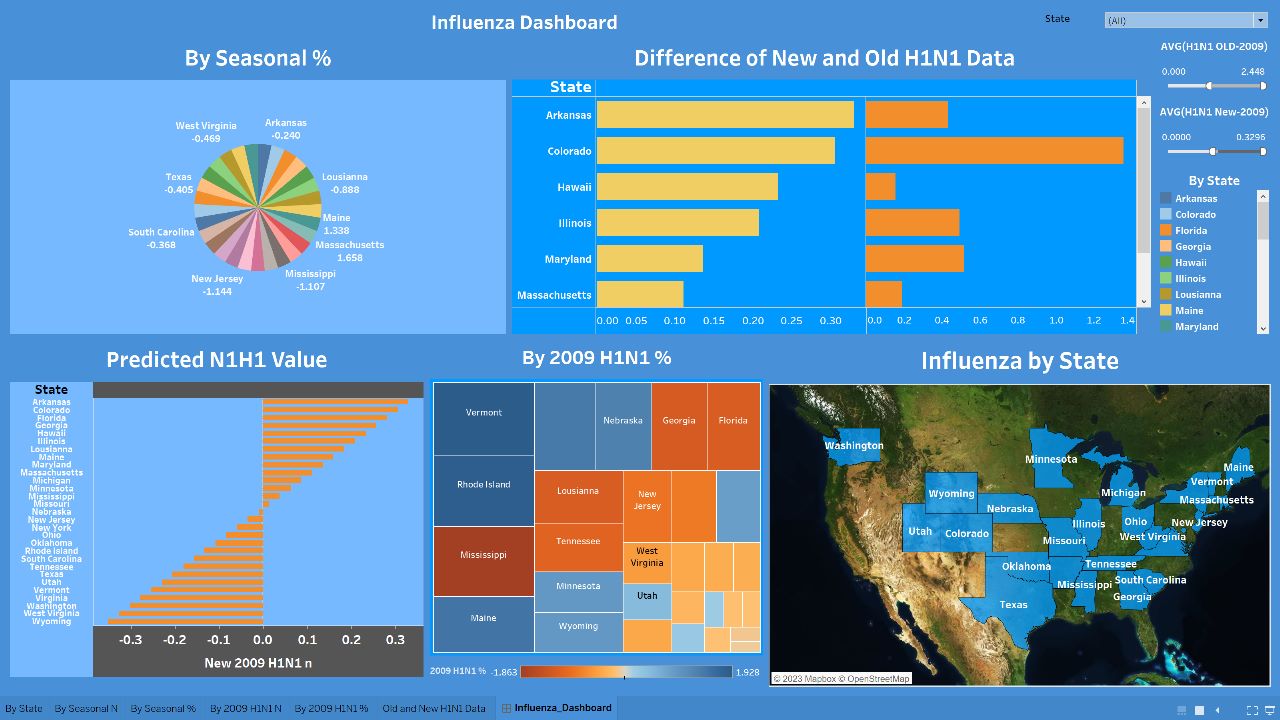


The user can download the previous uploaded data files from this page but the user should be logged in to download the files.

Data Virtualization Dashboard on Tableau

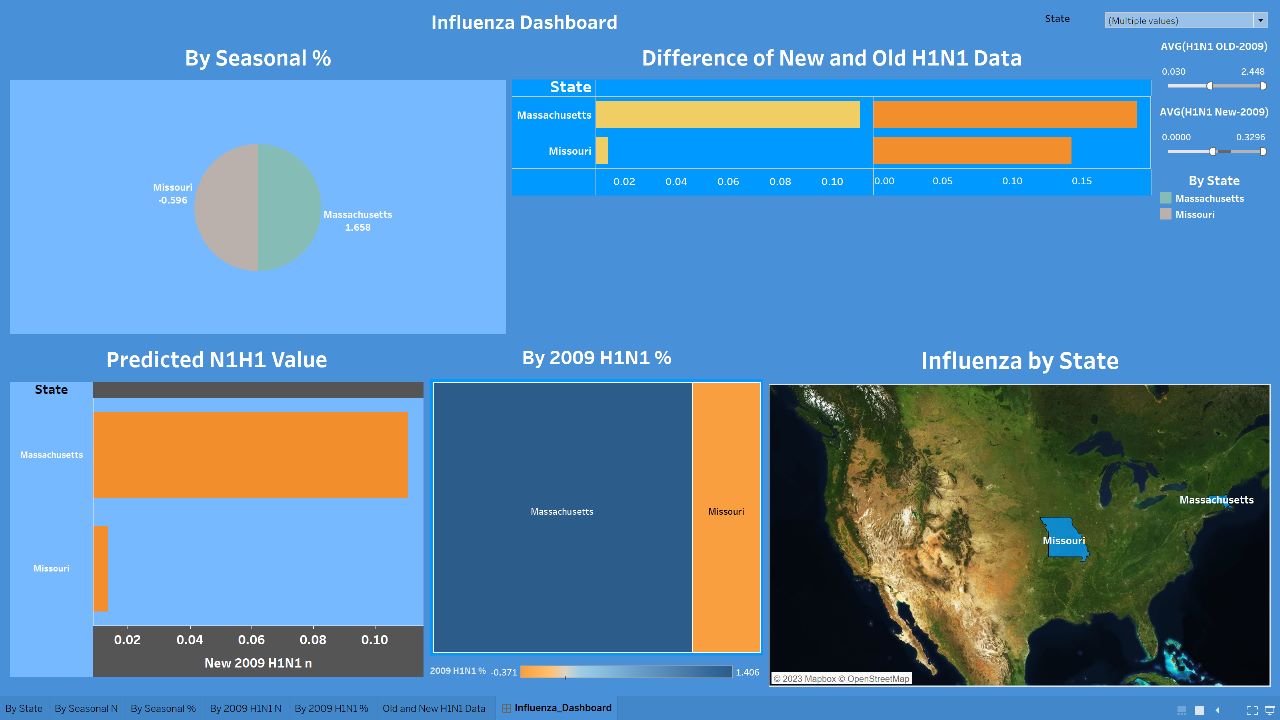
Data visualization is a way of presenting data in a visual form to make it easier to understand and analyze. Data dashboards are a summary of different, but related data sets, presented in a way that makes the related information easier to understand. We have used Tableau for the data virtualization.

## Main Dashboard



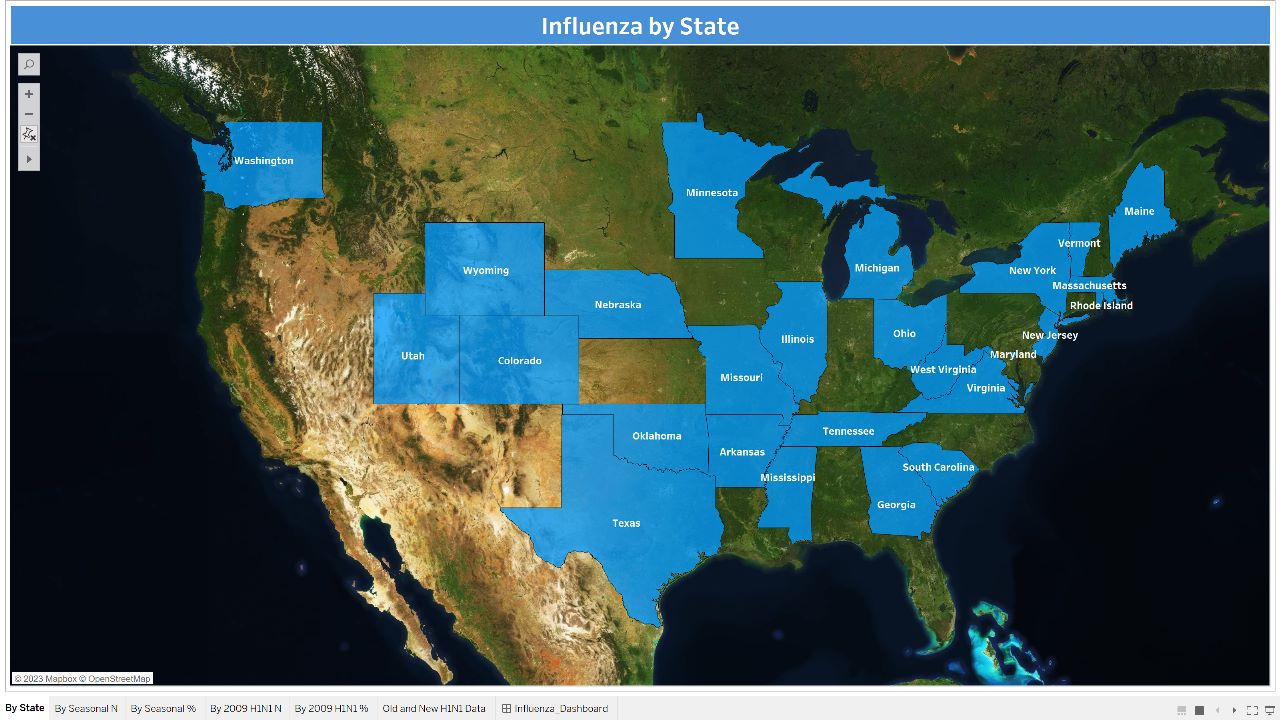
This is the dashboard designed on Tableau. This screen shows the multiple visualization angle for the predicted data. You can filter the data by state. The dashboard lets the user visualize the data of different fields so the user can visualize data according to their needs.

## Filtered Data by State



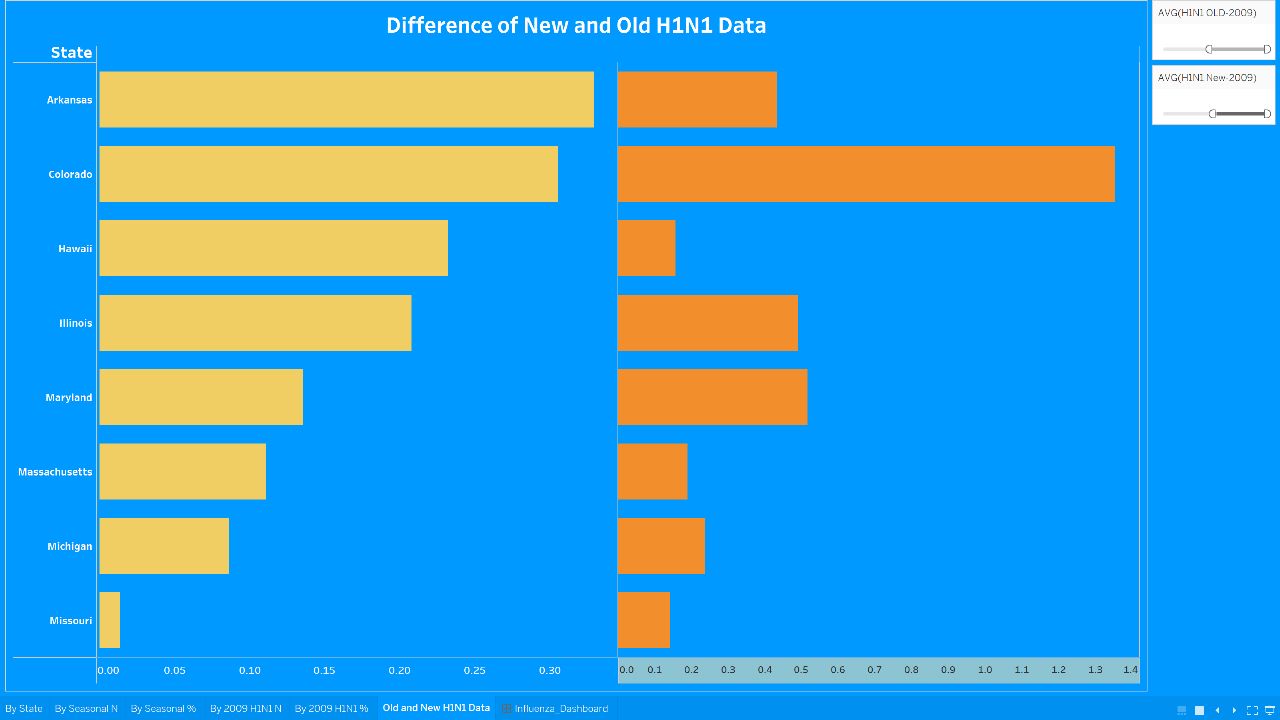
This data is filtered by state. The name of states are “Missouri” and “Massachusetts”.

## Map View



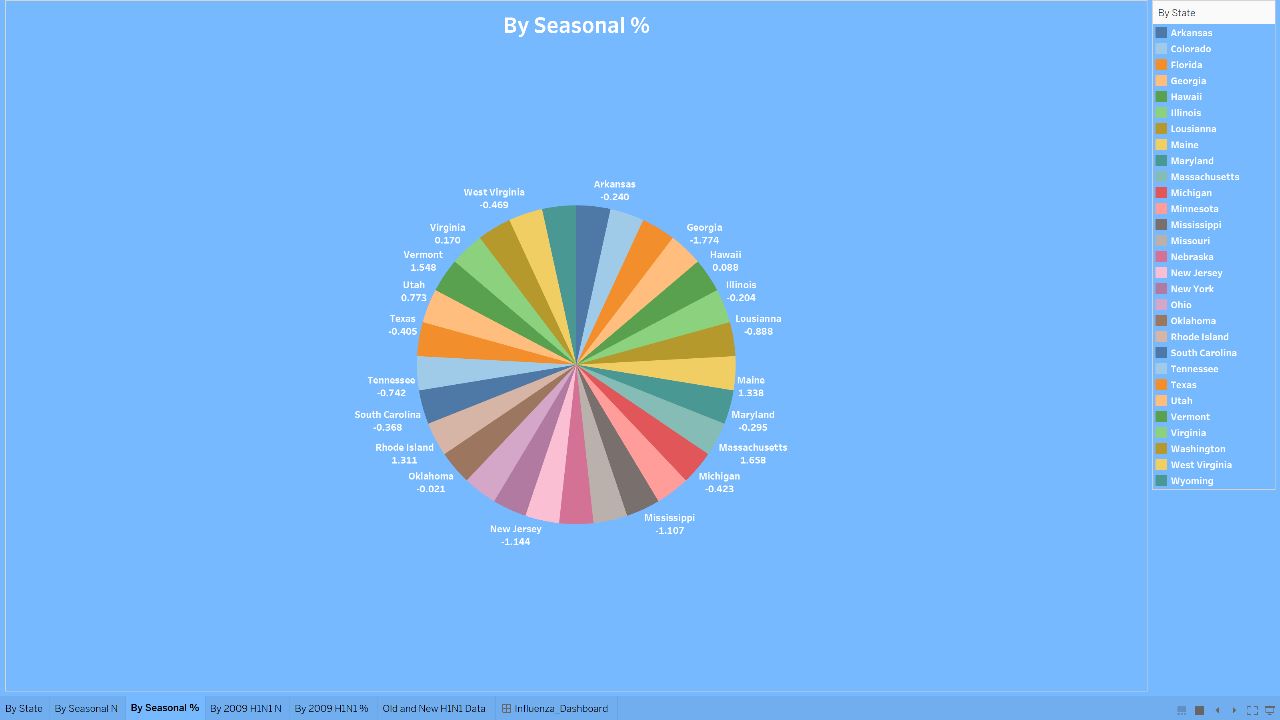
This is the overall view of data by the map.

## Comparison



This chart shows the comparison between the new and predicted values of H1N1. You can analyze by the chart that there are highly chances of Influenza virus to be spread.

## Values by Seasonal Percentage



This Pie chart visualizes the data by the seasonal percentages. The data can be filtered by the names of state.

# Dashboard Link:

Here is the link of dashboard

# Conclusion

In conclusion, the development of 'ViroShield' marks a significant stride towards enhancing our understanding of health issues and mortality causes. Through the adept utilization of Big Data techniques, this application has demonstrated its potential to predict health outcomes with a level of precision and accuracy previously unattainable. By harnessing the power of large and diverse datasets, 'ViroShield' efficiently processes, integrates, and analyzes information, leading to insightful predictions related to influenza and patient health.

The incorporation of Data Visualization techniques, particularly using tools like Tableau and Python Programming, has amplified the impact of the application. The provision of graphical representations through bar-graph charts and dashboard displays offers users a comprehensive view of predicted values based on meticulously trained datasets. This not only facilitates a clearer comprehension of health status but also empowers users to make informed decisions regarding their well-being.

As the field of healthcare continues to advance, 'ViroShield' serves as a testament to the potential of technology to revolutionize disease prediction and prevention. The convergence of Big Data analytics and Data Visualization has enabled this application to bridge the gap between raw data and actionable insights, fostering a proactive approach to health management.

In essence, the successful completion of the 'ViroShield' project underscores the transformative impact of interdisciplinary collaboration, innovation, and technology in the realm of healthcare. As we move forward, the lessons learned from this endeavor can guide future efforts to harness data-driven solutions for addressing health challenges, ultimately leading to improved outcomes and quality of life for individuals across the globe.