**Exploratory Data Analysis Report**

**Understanding Predictors of H1N1 and Seasonal Flu Vaccination**

**By**

Katta Manasa 20221CSE0685

Harshavardhan Kurthkoti 20221IST0101

Keerthi M Naik 20221IST0103

Aashna K Nishad 20221IST0073

Department of Computer Science Engineering

Presidency Univeristy

Itgalpura Rajanakunte,Yehlanka

Bengaluru – 560093 , Ph No – 080 23093500

February 2024

**Table of Contents**

Page 1: Introduction

* Brief Introduction to the Task and Objectives
* Significance of Understanding Factors Influencing H1N1 and Seasonal Flu Vaccination
* Highlight of Dataset's Source (2009 National H1N1 Flu Survey by CDC)

Page 2,3: Exploratory Data Analysis (Part 1)

* Importance of Exploratory Data Analysis (EDA)
* Summary Statistics of the Dataset
* Distribution Graphs of H1N1 Vaccine and Seasonal Vaccine
* Insights Gained from Distribution Graphs

Page 5: Exploratory Data Analysis (Part )

* Exploration of Missing Values and Strategies for Handling Missing Values

Page 6: Exploratory Data Analysis (Part 3)

* Display of the Correlation Matrix
* Discussion of Strong Correlations and Their Potential Impact
* Outlier Detection Methods and Identification of Outliers
* Additional Insights and Observations from EDA

Page 7: Exploratory Data Analysis (Part 4)

* Expansion on Additional Insights and Observations
* Context or Explanations for Anomalies or Interesting Patterns
* Discussion on How Insights May Inform Subsequent Analysis or Model Building

Page 8: Model Training and Evaluation

* Overview of Machine Learning Model Training Process
* Explanation of Evaluation Metrics (ROC AUC Score)
* Showcase of Performance of Trained Models
* Comparison of Different Models if Applicable

Page 9: Conclusion

* Summary of Key Findings from EDA and Model Training
* **Introduction:**

The AnalyticaX Data Science Challenge is a comprehensive project aimed at leveraging data science techniques to address a critical challenge. The project revolves around predicting the likelihood of individuals getting vaccinated for H1N1 and the seasonal flu. This challenge involves a multilabel classification problem, where the goal is to provide probabilities for each individual receiving both vaccines

**Task and objectives:**

"The task involves analyzing data from the 2009 National H1N1 Flu Survey conducted by the CDC to understand the factors influencing individuals' likelihood of receiving the H1N1 and seasonal flu vaccines. The primary objective is to identify key predictors of vaccine uptake to inform public health strategies and vaccination campaigns. Through exploratory data analysis and predictive modeling, we aim to uncover insights that can aid in promoting vaccine acceptance and uptake, ultimately contributing to improved public health outcomes."

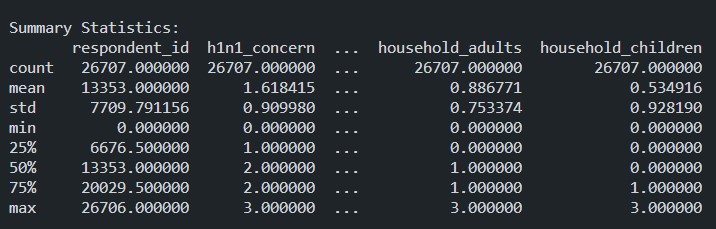
**Significance and Dataset source:**

"Understanding the factors influencing H1N1 and seasonal flu vaccination is crucial for effective public health policy and intervention strategies. By identifying predictors of vaccine uptake, we can better target at-risk populations, tailor communication and education efforts, and allocate resources efficiently to maximize vaccination coverage. This not only helps prevent the spread of influenza viruses but also reduces the burden on healthcare systems and minimizes the impact of seasonal epidemics and pandemics.

The dataset used for this analysis originates from the 2009 National H1N1 Flu Survey conducted by the Centers for Disease Control and Prevention (CDC). This comprehensive survey provides valuable insights into individuals' attitudes, behaviors, and demographics related to H1N1 and seasonal flu vaccination during the 2009 flu pandemic, offering a rich source of data for exploring vaccination trends and factors influencing vaccine uptake."

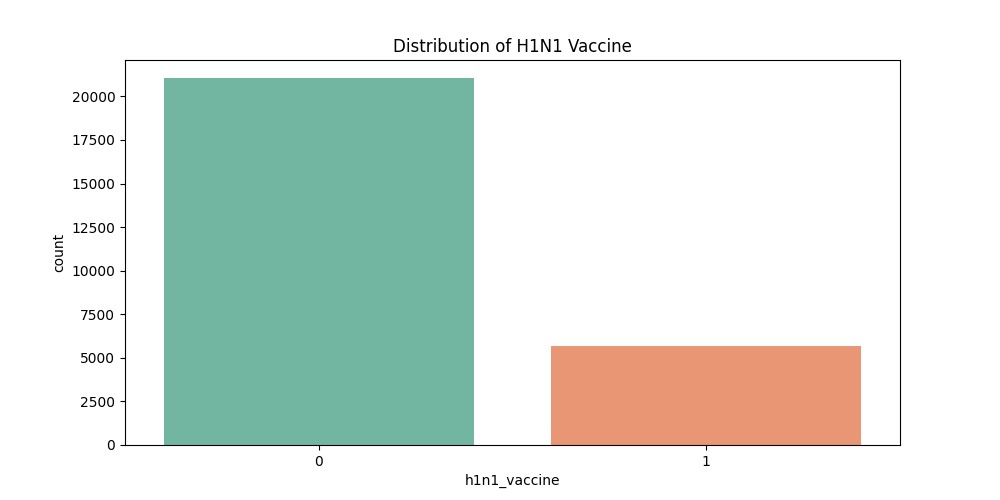
**Exploratory Data Analysis (EDA)** is essential for understanding a dataset's structure, patterns, and relationships. It uncovers insights, identifies trends, and informs further analysis. EDA leads to more meaningful interpretations and better-informed decisions by providing a deeper understanding of the data.

**Summary statistics** give a quick overview of important dataset features like average, middle value, variability, and extreme values. They help spot outliers and understand data distribution.

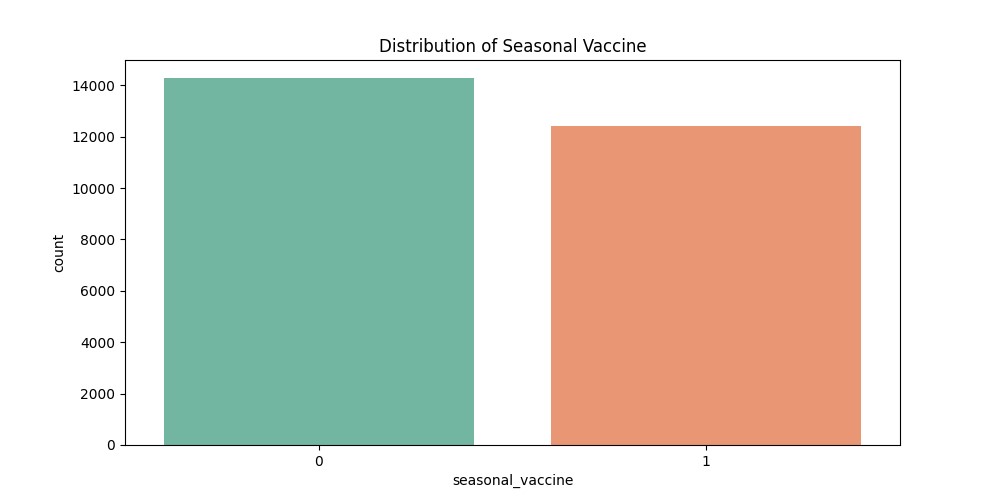


**Distribution graphs** For H1N1 and seasonal vaccine variables, these graphs reveal vaccination patterns among survey respondents. They help identify trends, assess vaccination prevalence, and spot potential class imbalances.

**Distribution of H1N1 Vaccine:**



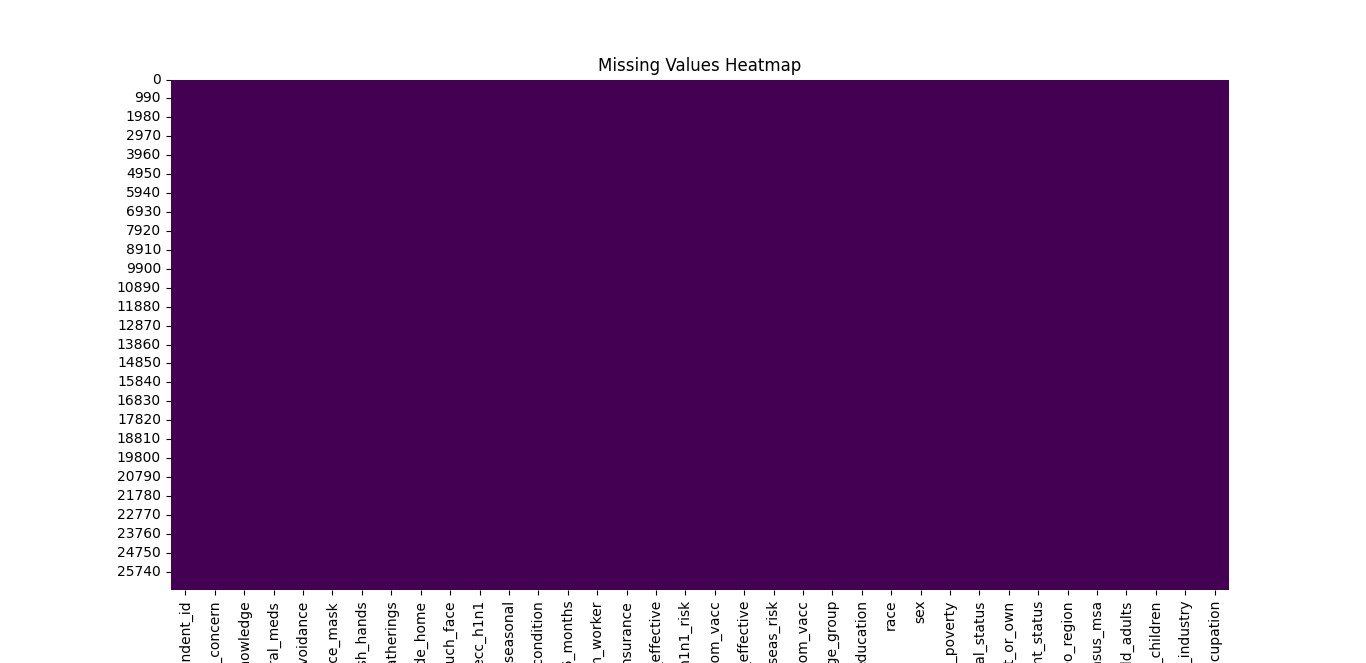
**Distribution of Seasonal Vaccine:**

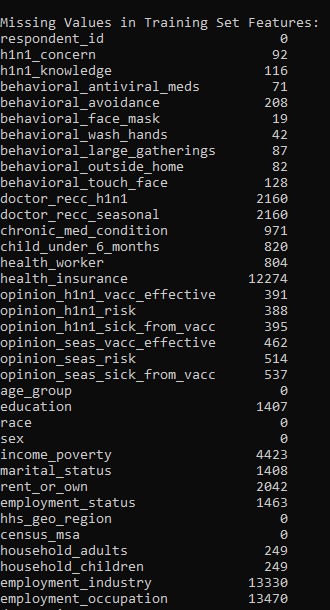


**Insights** from these distributions highlight vaccine uptake rates, class imbalances, and disparities among demographic groups. This information guides targeted interventions and public health initiatives to improve vaccine acceptance and uptake.

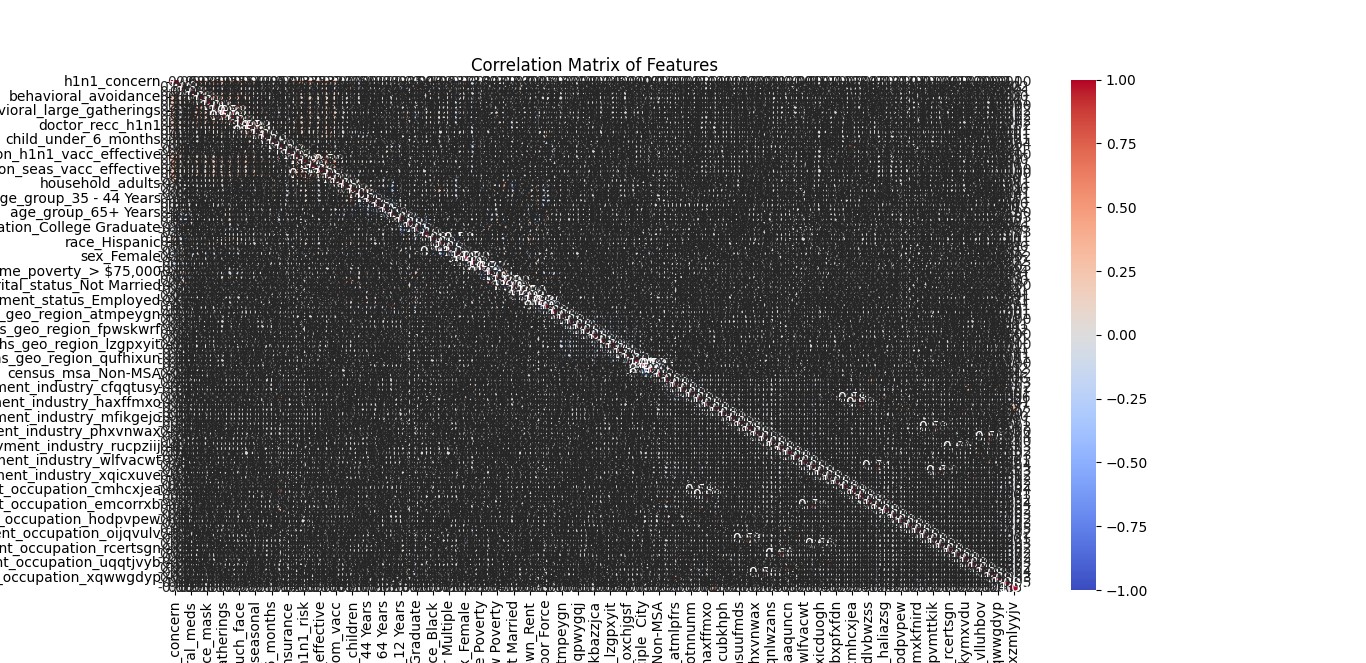
**Missing Values** - Certain features, such as "respondent\_id," "age\_group," "race," "sex," "hhs\_geo\_region," and "census\_msa," exhibit either no missing values or only a small number of missing values. These features are relatively complete and can be utilized directly in the analysis without the need for imputation or deletion.

**Missing Values Heatmap**

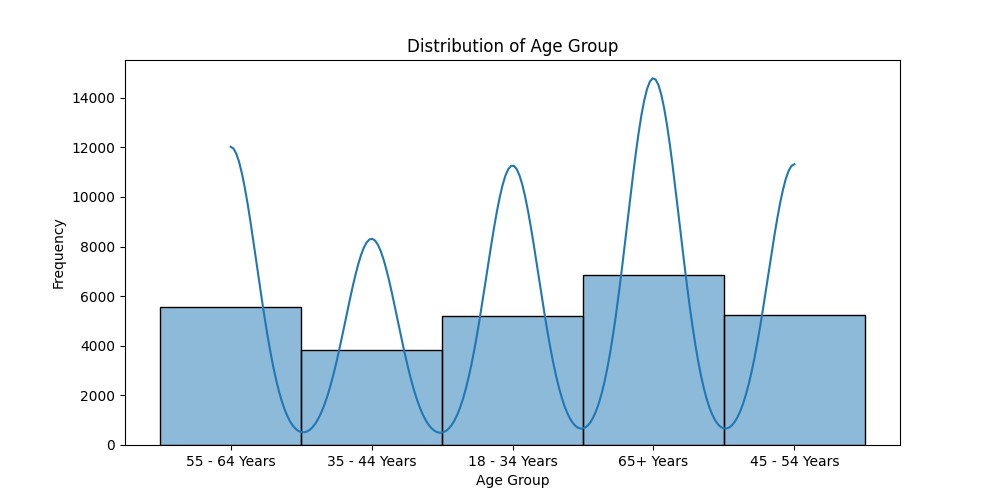




**Correlation matrix of features** provides insights into the linear relationships between numerical features in the dataset. Strong correlations, indicated by correlation coefficients close to 1 or -1, can have a significant impact on the analysis.



Analyzing the **Distribution of age groups** alongside vaccine uptake data provides valuable context for understanding the factors influencing H1N1 and seasonal flu vaccination among different age cohorts.



.