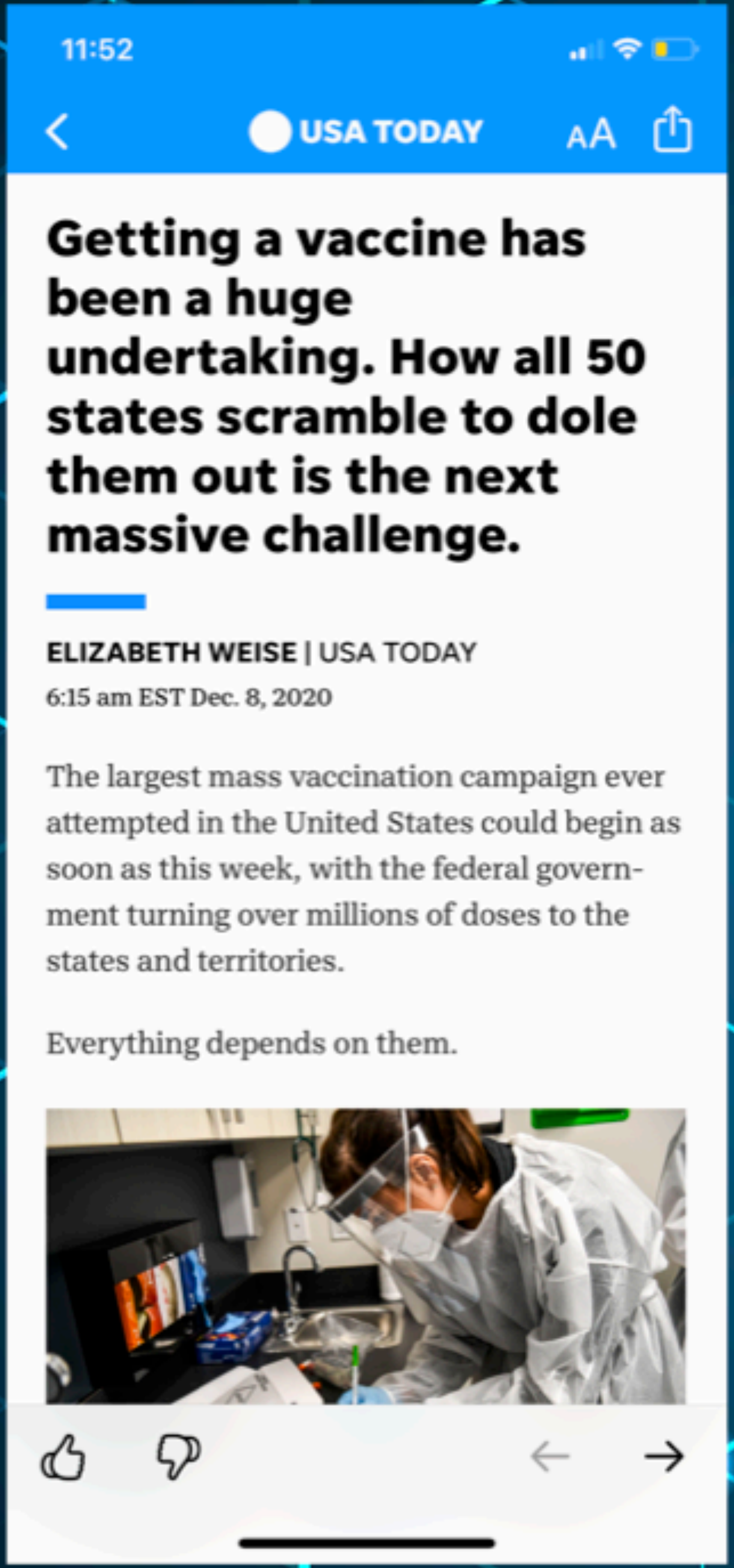
H1N1 Vaccine Prediction

Group 10

Allen Lee
Eric Chen
Hyoungmin (Stella) Lee
Marianna Carini
Smitha Kannanaikkal



11:52

USA TODAY

Getting a vaccine has been a huge undertaking. How all 50 states scramble to dole them out is the next massive challenge.

ELIZABETH WEISE | USA TODAY

6:15 am EST Dec. 8, 2020

The largest mass vaccination campaign ever attempted in the United States could begin as soon as this week, with the federal government turning over millions of doses to the states and territories.

Everything depends on them.

Introduction

- Using National 2009 H1N1 Flu Survey to predict vaccine demand in the United States
- Focusing on quantity of production

Business Problem

Why is this important?

- Pharmaceutical companies need an estimate as to not overproduce or underproduce
- Would benefit agencies like the World Health Organization and healthcare providers to plan distribution and receival of large quantities of vaccines

Methodology

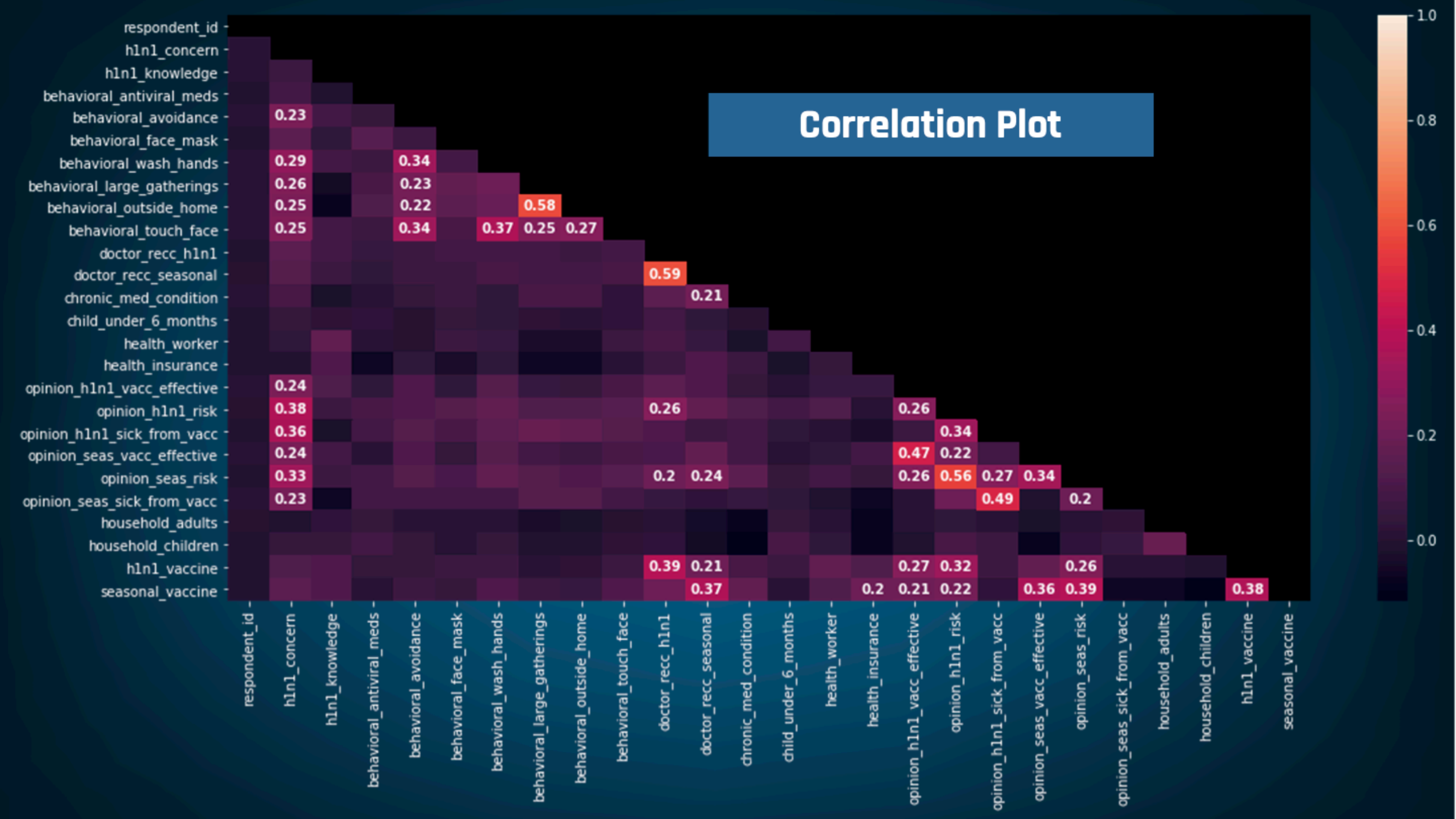
Dataset Details

DRIVENDATA

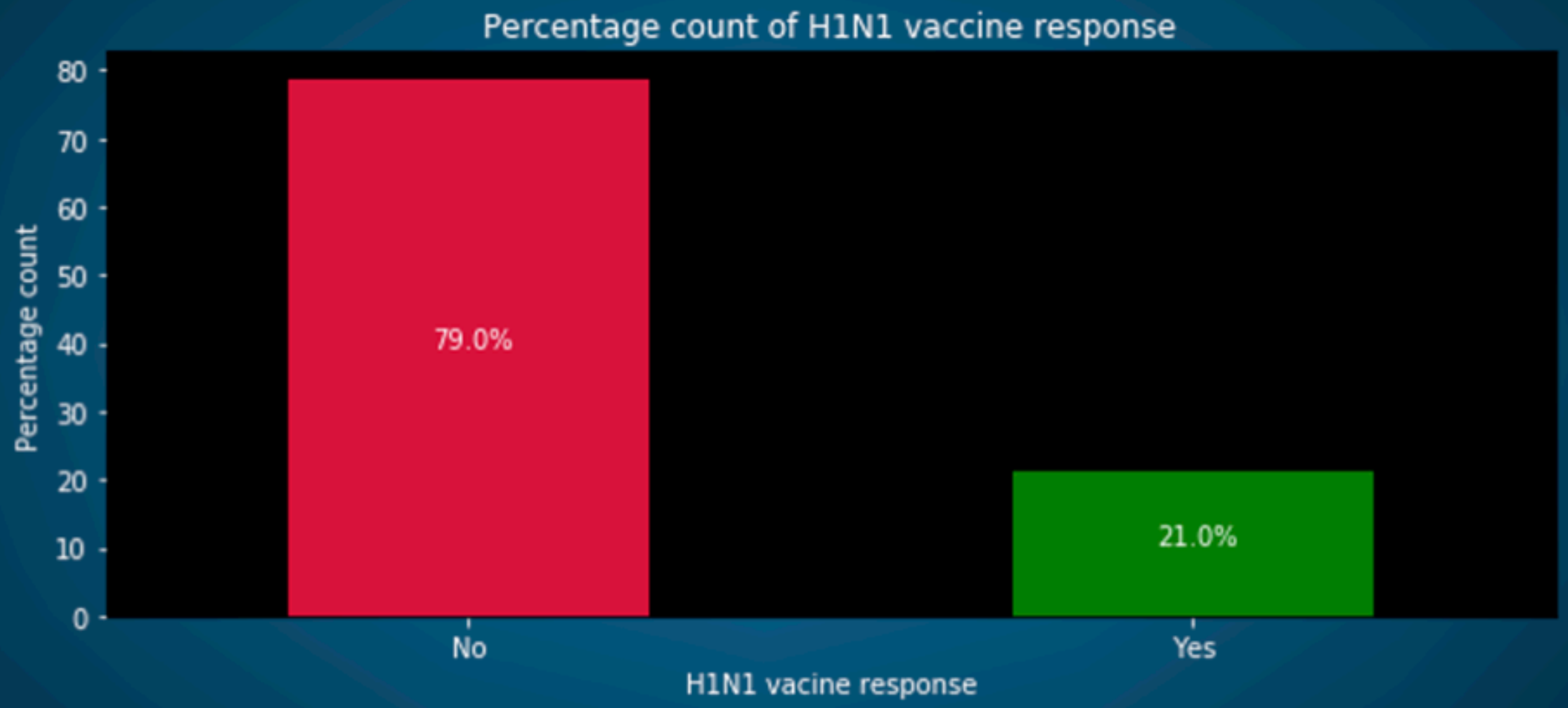
Number of attributes: 38 | **Number of Instances: 26706**

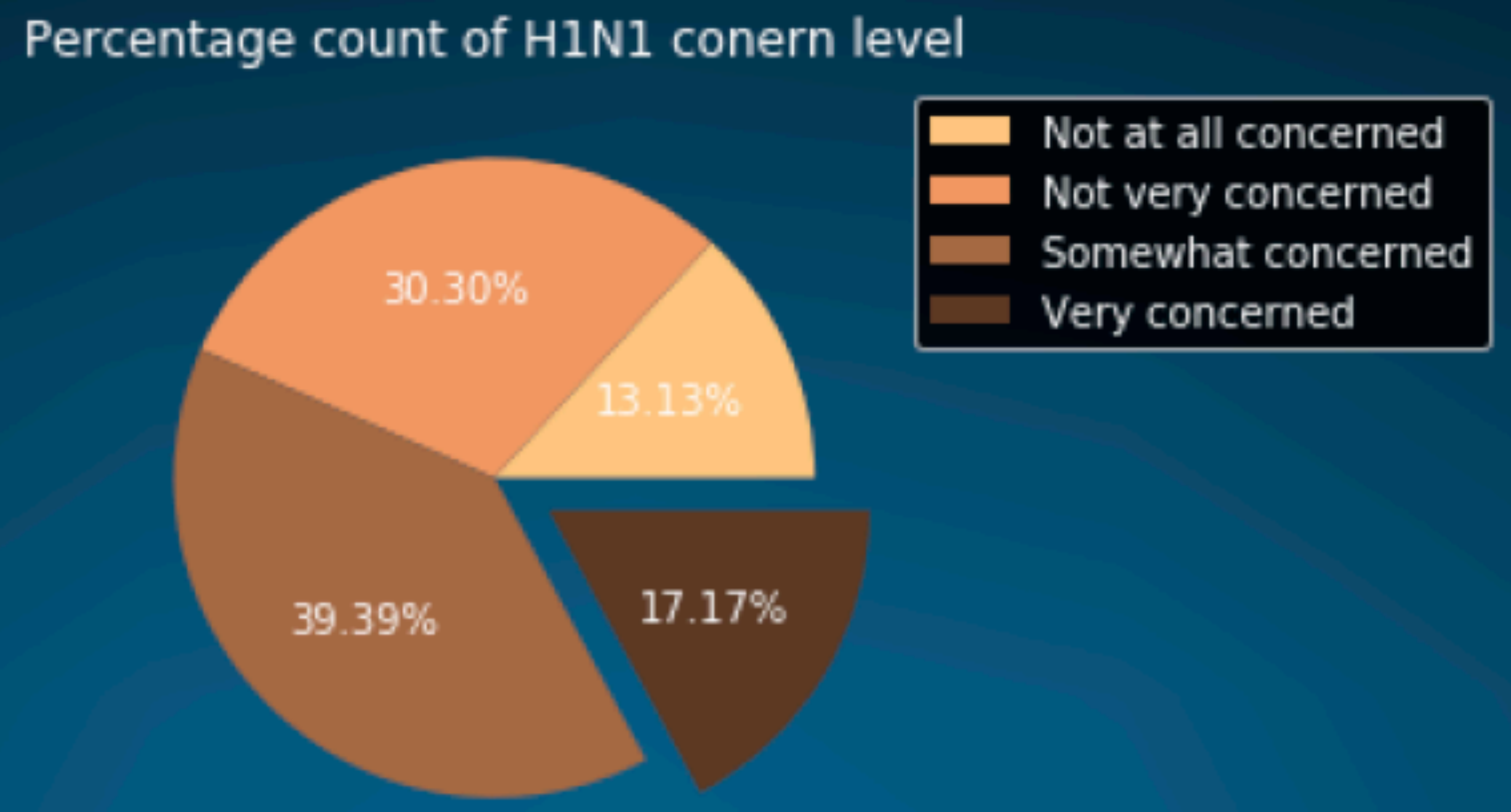
Behavioural	Opinion	Demographic	Others
antiviral_meds	h1n1_vacc_effctive	age_group	h1n1_knowledge
avoidance	h1n1_risk	education	doctor_recc_h1n1
face_mask	h1n1_sick_from_vacc	race	doctor_recc_seasonal
wash_hands	seas_vacc_effective	sex	chronic_med_condition
large_gatherings	seas_risk	income_poverty	child_under_6_months
outside_home	seas_sick_from_vacc	marital_status	health_worker
touch_face	h1n1_concern	rent_or_own	health_insurance
		household_adults	h1n1_vaccine
		household_children	employment_industry
		employment_status	employment_occuption
		census_msa	hhs_geo_region

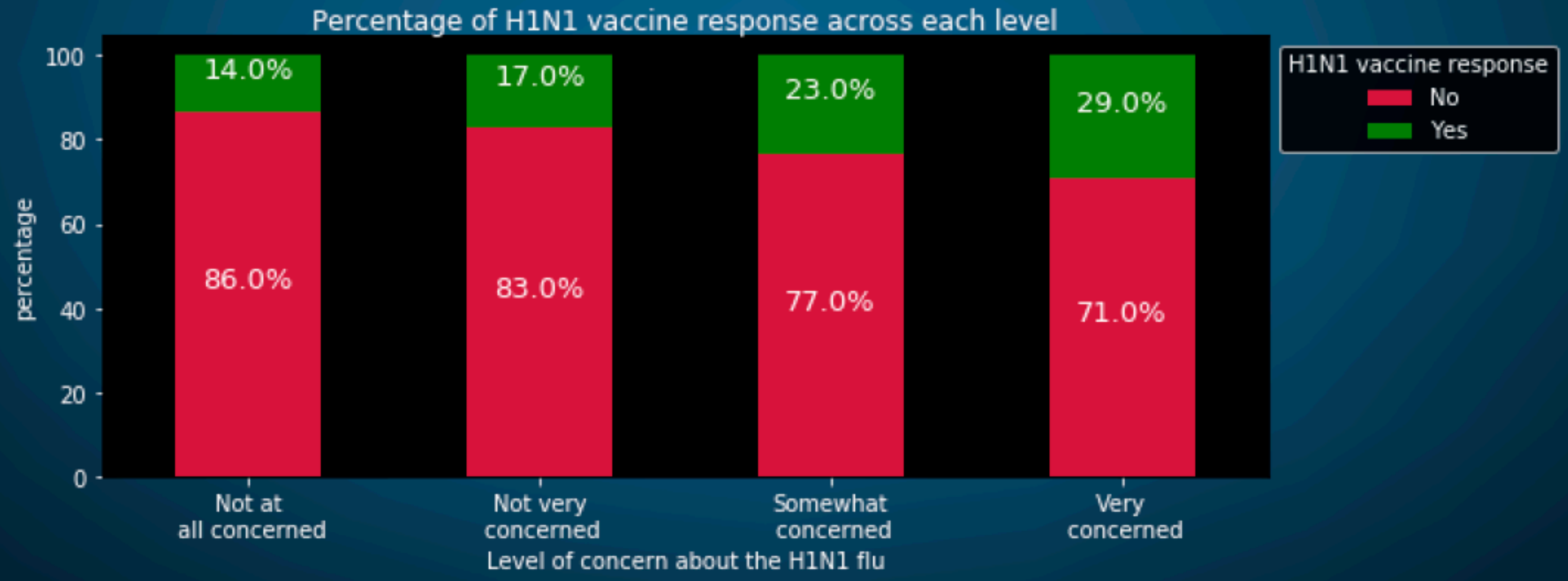
Exploratory Data Analysis


Correlation Plot
respondent_id
h1n1_concern
h1n1_knowledge
behavioral_antiviral_meds
behavioral_avoidance
behavioral_face_mask
behavioral_wash_hands
behavioral_large_gatherings
behavioral_outside_home
behavioral_touch_face
doctor_recc_h1n1
doctor_recc_seasonal
chronic_med_condition
child_under_6_months
health_worker
health_insurance
opinion_h1n1_vacc_effective
opinion_h1n1_risk
opinion_h1n1_sick_from_vacc
opinion_seas_vacc_effective
opinion_seas_risk
opinion_seas_sick_from_vacc
household_adults
household_children
h1n1_vaccine
seasonal_vaccine
1.0
0.8
0.6
0.4
0.2
0.0

Class Imbalance

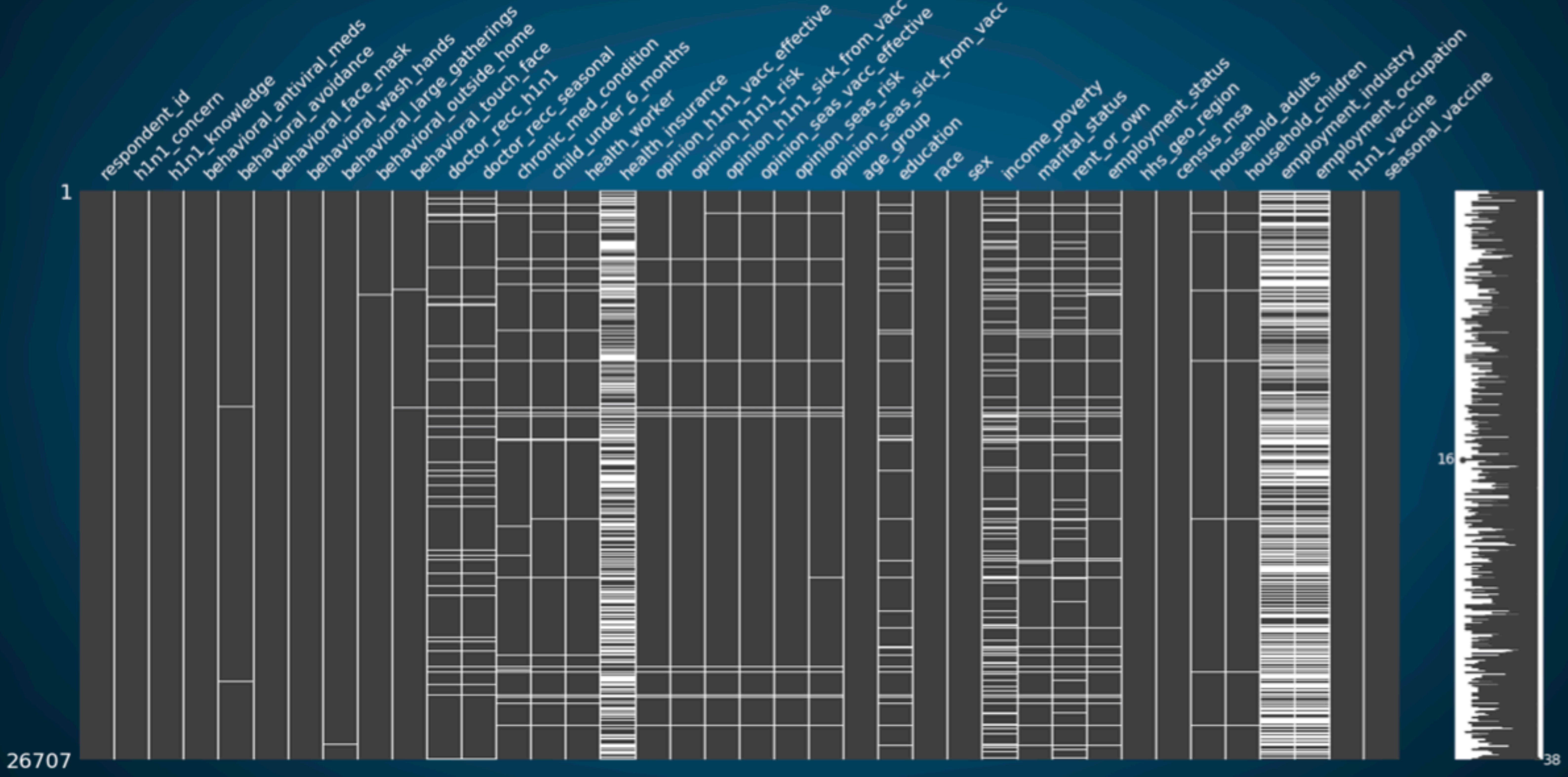



Percentage count of H1N1 conern level
Not at all concerned
Not very concerned
Somewhat concerned
Very concerned
30.30%
13.13%
39.39%
17.17%


Percentage of H1N1 vaccine response across each level
H1N1 vaccine response
No
Yes
percentage
0
20
40
60
80
100
14.0%
86.0%
17.0%
83.0%
23.0%
77.0%
29.0%
71.0%
Not at all concerned
Not very concerned
Somewhat concerned
Very concerned
Level of concern about the H1N1 flu

Data Cleaning

Missing Value in Data



Missing Value Imputation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naïve Bayes

Imputed attribute by prediction

Null Pattern

Imputed attribute based on the Null value pattern

Median/Mode

Imputed attribute based on central measure

Model Building

Machine Learning Models

Model 1

Naïve Bayes

Model 2

Decision Tree

Model 3

Random Forest

Several versions of each model were built

- Forward feature selection
- Oversampling
- 10-fold cross validation

Performance measure

Precision Recall Fscore

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Fscore} = \frac{2*Precision*Recall}{Precision+Recall}$$

Model Performance

Naïve Bayes	Accuracy	Class 1 Precision	Class 1 Recall	Class 1 F1-Score	Class 1 ROC_AUC
Benchmark	79.99%	0.526	0.633	0.574	0.828
Forward Selection	84.35%	0.692	0.474	0.563	0.849
Oversampling	73.97%	0.740	0.740	0.740	0.822
Oversampling Forward Select	78.25%	0.793	0.769	0.781	0.856

Decision Tree	Accuracy	Class 1 Precision	Class 1 Recall	Class 1 F1-Score	Class 1 ROC_AUC
Benchmark	84.31%	0.670	0.521	0.586	0.798
Forward Selection	85.01%	0.693	0.533	0.602	0.839
Oversampling	83.83%	0.827	0.855	0.841	0.863
Oversampling Forward Select	83.91%	0.830	0.825	0.861	0.860

Random Forest	Accuracy	Class 1 Precision	Class 1 Recall	Class 1 F1-Score	Class 1 ROC_AUC
Benchmark	84.69%	0.721	0.460	0.562	0.868
Forward Selection	84.78%	0.671	0.554	0.607	0.637
Oversampling	91.23%	0.893	0.937	0.914	0.974
Oversampling Forward Select	91.48%	0.896	0.939	0.917	0.974

Model Evaluation

		Training Set Performance					Testing Set Performance				
Model	**Process**	**Accuracy**	**Class 1 Precision**	**Class 1 Recall**	**Class 1 F1-Score**	**Class 1 ROC_AUC**	**Accuracy**	**Class 1 Precision**	**Class 1 Recall**	**Class 1 F1-Score**	**Class 1 ROC_AUC**
Naïve Bayes	Oversampling Forward Select	79.25%	0.800	0.780	0.790	0.865	79.56%	0.519	0.777	0.612	0.868
Decision Tree	Oversampling	85.00%	0.836	0.871	0.853	0.870	77.70%	0.491	0.747	0.592	0.758
Random Forest	Oversampling Forward Select	91.74%	0.903	0.935	0.919	0.974	81.20%	0.552	0.713	0.622	0.864

Conclusion

- Naïve Bayes was the best performing model in terms of minimizing potential loss in sales/revenue

- Using the confusion matrix we estimate 107.1 million doses will be needed

Thank you

Do you have any questions?