Predicting the Likelihood of H1N1 Vaccination using Data Mining Methods:

Analyzing Behavioral and Demographic Data

Luke Awino, Roberto Cancel, Kevin Stewart

University of San Diego

ADS 502: Applied Data Mining

Dr. Ebrahim Tarshizi

8/16/2021

Abstract

Behavioral and demographic factors affect the likelihood of vaccination. The problem for public health practitioners is better understanding the socio-demographic features that most influence the likelihood of vaccination. This study analyzed 26,707 observations from the National 2009 H1N1 Flu Survey conducted by the National Center for Health Statistics to predict the likelihood of vaccination based on behavioral and demographics of the respondents. After cleaning and wrangling, we were left with 19,642 complete cases to use in our data. We deployed Logistic Regression, Naïve Bayes, and Random Forest algorithms to model the likelihood of vaccination. Since we were most interested in predicting True Positive predictions or accurate vaccination predictions to determine the features that most influenced respondents to get vaccinated, we used sensitivity/recall as our measure of success. Logistic Regression with all features and reduced features resulted in a 62.71% sensitivity. Random Forest and Random Forest using the Mean Decrease Gini resulted in a 58.28% sensitivity. Naïve Bayes method resulted in a 70.90% sensitivity, while the improved Naïve Bayes using the Laplacian and kernel method resulted in a 69.32% sensitivity. This study concluded that the Naïve Bayes method was the most precise model to predict the likelihood of vaccination using behavioral and demographic data. However, further evaluation for use in public health communication efforts are needed as we continue living in this ever-evolving world with frequent viral plagues.

Keywords: H1N1, vaccine prediction, Logistic Regression, Naïve Bayes, Random Forest, Behavioral data, Demographic data

Table of Contents

Introduction

Nearly a decade ago, public health professionals battled the influenza A virus subtype H1N1 (H1N1) global pandemic. H1N1 was a new influenza virus, giving health care experts an advantage in producing and distributing the H1N1 vaccine since they had years of experience with influenza vaccination. However, the public health effort to develop and distribute a safe and effective H1N1 vaccination campaign was plagued with challenges, including communication of vaccine availability and suggested participation, vaccine supply chain issues, and public concern regarding the safety and efficacy of the H1N1 vaccine. "These challenges eroded public trust in the H1N1 vaccination program: a November 2009 survey found that 54% of adults believed the federal government was doing a "poor" or "very poor" job at providing the country with an adequate supply of H1N1 vaccine" (Newport, 2021).

The objective of this study was to deploy data mining methods to determine our ability to predict the likelihood of a patient adopting the H1N1 vaccine using behavioral and socio-demographic data. The initial goal was to determine which method most accurately predicted H1N1 vaccine adoption. The secondary goal is to determine the most influential features in the likelihood of H1N1 vaccine adoption to help with current COVID19 public health communication and future vaccination efforts.

Methodology

Data Collection and Pre-processing

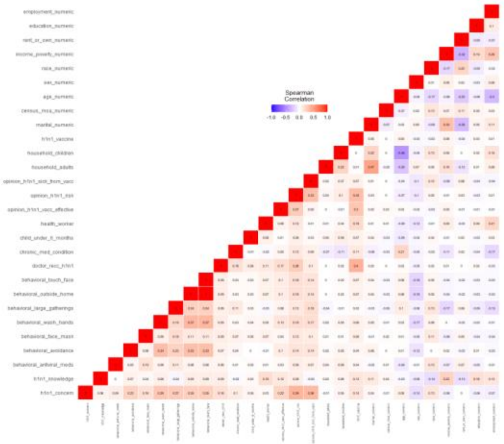
The study is one portion of DataDriven.com's Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines competition using data from the National 2009 H1N1 Flu Survey conducted by the National Center for Health Statistics. They surveyed 26,707 respondents on vaccination status and 35 features, including their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission.

We used the R programming language throughout this study. First, the training feature data set was merged with the target data set to create a comprehensive data set and subsequent model evaluation partitioning. The provided test set did not include target information due to the nature of the DataDriven competition. Once merged, a summary of the features in the dataset revealed extensive missing values for features throughout the dataset. Initially, we attempted imputation by mode for each feature; however, we decided that a complete case review would reduce any potential imputation bias, reducing our observations to 19,642. Due to our objective, we also removed seasonal flu-associated variables. Finally, since most of our features were binary responses, we evaluated the distribution of our nine categorical variables: age_group, education, race, sex, income_poverty, marital_status, rent_or_own, employment_status, and census_msa.

We re-expressed these categorical variables as numerical for use in subsequent modeling. Since our data was socially focused, we determined it was best to use all features after rationalizing for multicollinearity. Figure 1 shows the Spearman correlation of our features. We used the Spearman correlation because all of our features were binary or ordinal, and Spearman correlations exist to measure the association of variables rather than their distance. For example, we see multicollinearity exists between behavioral_touch_face and behavioral_outdoor. We decided to remove behavioral_touch_face to ensure no multicollinearity exists before modeling.

Figure 1

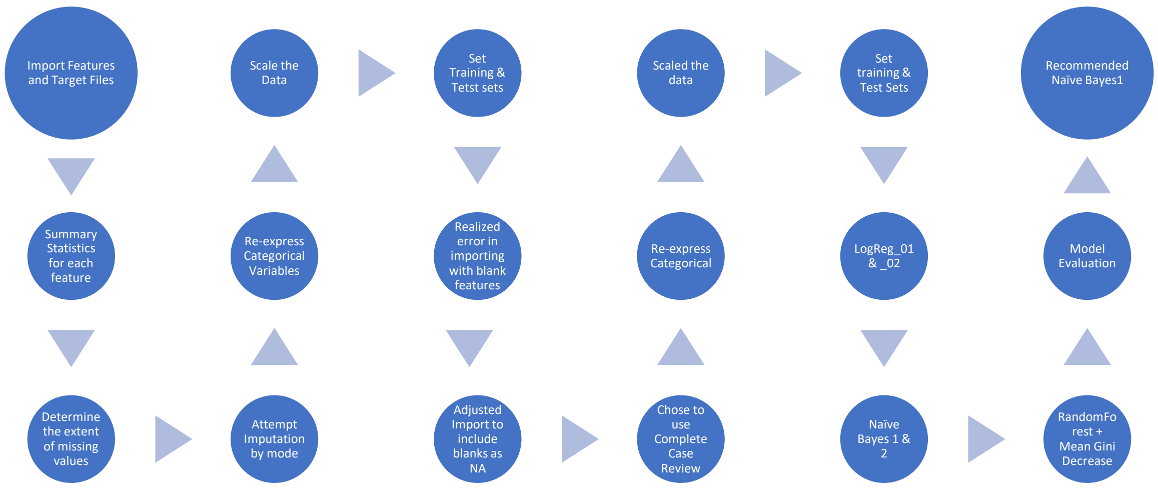
Spearman correlation of features



All features were then scaled (min-max), and we partitioned the data set into a 75% training set and 25% test set. Following partitioning, the training set was balanced by randomly oversampling the h1n1_vaccine True/1 records by 7,913 observations. Finally, we attempted to cross-validate the partitioning, but we could not use the Welch Two-Sample T-test since our features were binary and ordinal. Figure 2 depicts our general workflow.

Figure 2

Project Workflow


Import Features and Target Files
Summary Statistics for each feature
Determine the extent of missing values
Attempt Imputation by mode
Re-express Categorical Variables
Scale the Data
Set Training & Tetst sets
Realized error in importing with blank features
Adjusted Import to include blanks as NA
Chose to use Complete Case Review
Re-express Categorical
Scaled the data
Set training & Test Sets
LogReg_01 & _02
Naïve Bayes 1 & 2
RandomForest + Mean Gini Decrease
Model Evaluation
Recommended Naïve Bayes1

Modeling

Logistic Regression (Roberto Cancel)

Since we were predicting a binary target variable, we chose Logistic Regression as one of our models. We ran two iterations, initially with all features (LogReg_01) and a second with primarily statistically significant variables (LogReg_02) with some essential demographic features retained regardless of significance.

LogReg_01's baseline model contained many statistically insignificant variables, with the five most impactful variables on predicting the likelihood of vaccination being opinion_h1n1_vacc_effective (2.620), opinion_h1n1_sick_from_vacc (2.620), opinion_h1n1_risk, doctor_recc_h1n1 (1.700), age_numeric (0.463), and h1n1_knowledge (0.343). More detailed information on the remaining variables is shown in Table 1. When evaluated LogReg_01 against the test data set, the sensitivity for the logistic regression model was determined to be 62.71%.

Logreg_02's baseline showed a new set of statistically insignificant variables with the five most impactful variables on predicting the likelihood of vaccination being opinion_h1n1_vacc_effective (2.620), opinion h1n1_risk (1.879), doctor_recc_h1n1 (1.699), health_worker (1.007), and age_numeric (0.4839). when evaluated against the test data set, the sensitivity for the logistic regression model was also determined to be 62.71% - suggesting the previous removal did not impact sensitivity.

Since census_msa_numeric, race_numeric, and employment_numeric features are considered essential to include in socio-demographic studies, they were left in the model, resulting in our final descriptive logistic regression equation:

$$\hat{p}(\mathrm{h1n1\_vaccine}) = \frac{\exp(-4.14 - 0.278(concern) + 0.344(knowledge) + 0.135(antiviral) - 0.094(avoidance) + 0.141(mask) + 0.094(handwash) - 0.250(gatherings) + 1.699(doctorrecc) + 0.141(chonicmed) + 0.234(child < 6mons) + 1.007(healthworker) + 2.620(vacc\ eff) + 1.879(risk) - 0.124(children) + 0.143(marital) - 0.014(census) + 0.483(age) + 0.208(sex) + 0.104(race) + .268(income) + 0.15(education) - .111(education)}{1 + \exp(-4.14 - 0.278(concern) + 0.344(knowledge) + 0.135(antiviral) - 0.094(avoidance) + 0.141(mask) + 0.094(handwash) - 0.250(gatherings) + 1.699(doctorrecc) + 0.141(chonicmed) + 0.234(child < 6mons) + 1.007(healthworker) + 2.620(vacc\ eff) + 1.879(risk) - 0.124(children) + 0.143(marital) - 0.014(census) + 0.483(age) + 0.208(sex) + 0.104(race) + .268(income) + 0.15(education) - .111(education)}$$

(1)

Further evaluating the logistic regression output indicates that efforts to increase the likelihood should be directed towards public awareness of the safety, efficacy, and risk associated with the virus and vaccine with targeted campaigns for younger people and encouragement for doctors to recommend vaccine to their patients.

Table 1:

LogReg_01 Feature values w/significance

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.08998	0.11984	-34.13	< 2e-16
h1n1_concern	-0.26952	0.06635	-4.062	4.86E-05
h1n1_knowledge	0.34285	0.05889	5.822	5.82E-09
behavioral_antiviral_meds	0.13732	0.07469	1.839	0.065986
behavioral_avoidance	-0.08552	0.0418	-2.046	0.040758
behavioral_face_mask	0.14651	0.06571	2.23	0.025781
behavioral_wash_hands	0.10647	0.05173	2.058	0.039597
behavioral_large_gatherings	-0.2434	0.03795	-6.414	1.42E-10
behavioral_outside_home	-0.04235	0.03981	-1.064	0.287455
doctor_recc_h1n1	1.70075	0.03759	45.247	< 2e-16
chronic_med_condition	0.14094	0.03731	3.777	0.000158
child_under_6_months	0.23488	0.05871	4.001	6.31E-05
health_worker	1.01	0.05039	20.044	< 2e-16
opinion_h1n1_vacc_effective	2.62019	0.08218	31.885	< 2e-16
opinion_h1n1_risk	1.88648	0.05679	33.218	< 2e-16
opinion_h1n1_sick_from_vacc	-0.02167	0.05353	-0.405	0.685618
household_adults	-0.08949	0.07521	-1.19	0.234103
household_children	-0.12649	0.06216	-2.035	0.041868
marital_numeric	0.15975	0.03972	4.022	5.77E-05
census_msa_numeric	-0.01302	0.04466	-0.292	0.770595
age_numeric	0.46256	0.05945	7.781	7.21E-15
sex_numeric	0.20681	0.03482	5.939	2.87E-09
race_numeric	0.11283	0.06129	1.841	0.065637
income_poverty_numeric	0.25982	0.06312	4.116	3.85E-05
rent_or_own_numeric	-0.01659	0.04484	-0.37	0.711478
education_numeric	0.15586	0.05832	2.672	0.007533
employment_numeric	-0.11605	0.06079	-1.909	0.056247

Table 2:

LogReg_02 Feature values w/significance

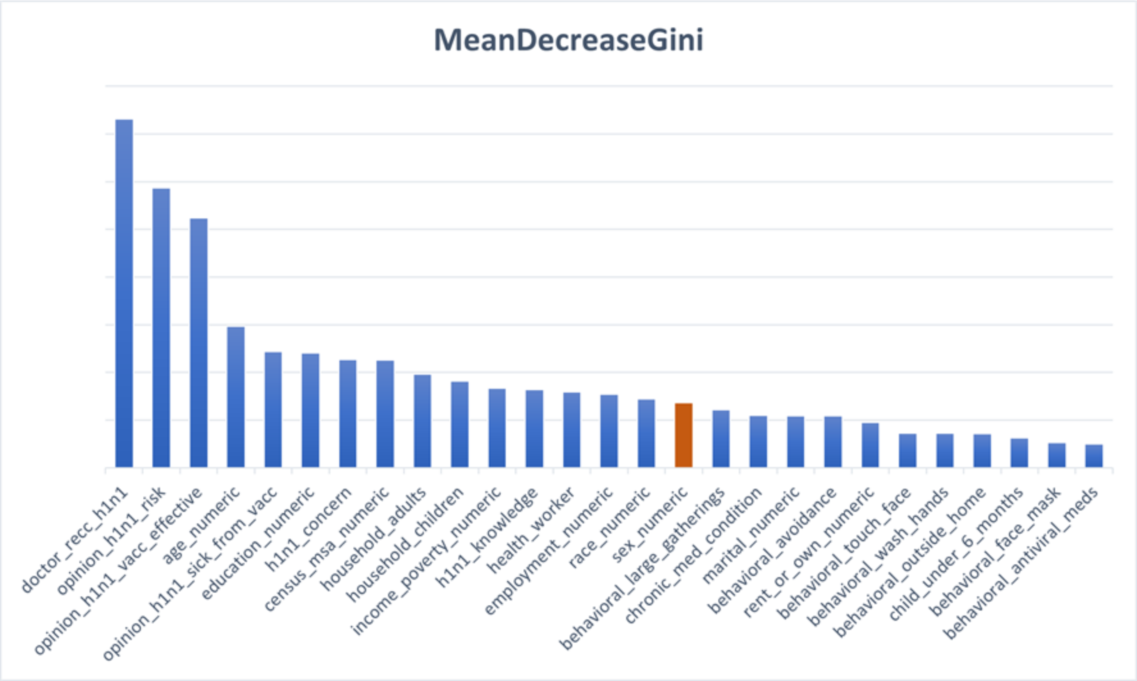
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.14193	0.11183	-37.038	< 2e-16
h1n1_concern	-0.27846	0.06457	-4.313	1.61E-05
h1n1_knowledge	0.34402	0.05869	5.862	4.58E-09
behavioral_antiviral_meds	0.13516	0.07466	1.81	0.070243
behavioral_avoidance	-0.09393	0.04107	-2.287	0.022181
behavioral_face_mask	0.14126	0.06558	2.154	0.031251
behavioral_wash_hands	0.09426	0.05057	1.864	0.062347
behavioral_large_gatherings	-0.24933	0.03754	-6.642	3.09E-11
doctor_recc_h1n1	1.69929	0.03756	45.242	< 2e-16
chronic_med_condition	0.14119	0.03727	3.788	0.000152
child_under_6_months	0.23436	0.0587	3.992	6.54E-05
health_worker	1.007	0.05023	20.049	< 2e-16
opinion_h1n1_vacc_effective	2.61975	0.08202	31.942	< 2e-16
opinion_h1n1_risk	1.8786	0.05533	33.951	< 2e-16
household_children	-0.12357	0.06196	-1.994	0.046126
marital_numeric	0.14307	0.03638	3.933	8.40E-05
census_msa_numeric	-0.01374	0.04445	-0.309	0.757292
age_numeric	0.48349	0.05569	8.681	< 2e-16
sex_numeric	0.20842	0.03465	6.015	1.80E-09
race_numeric	0.10442	0.06065	1.722	0.085123
income_poverty_numeric	0.26739	0.06152	4.347	1.38E-05
education_numeric	0.15843	0.05828	2.719	0.006555
employment_numeric	-0.11123	0.06058	-1.836	0.066363

Random Forest (Luke Awino)

Random Forest was run with 100 trees. The number of variables at each split was five; 27 variables were used in the dataset. Random Forests are robust to overfitting by considering strong and weak attributes and aggregating the predictions (Tan et al., 2019). The out-of-bag error rate is 6.85%. For the Random Forest model, variables with the highest 60 percent scores were picked in figure (2) below using the Mean Decrease Gini. And the Random Forest was rerun using the new variables, and the new error rate was returned. "The mean decrease in Gini coefficient measures how each variable contributes to the homogeneity of the nodes and leaves in the resulting Random Forest. The higher the value of the mean decrease Gini score, the higher the importance of the variable in the model (Martinez-Taboado & Redondo, 2020).

Figure 3

Selected Variables from Mean Decrease Gini output



The out-of-bag error rate for the updated model was 8.59% showing a decrease in the model's accuracy when only 60% of the variables are used based on the Mean Decrease Gini. However, the sensitivity from the test model was 56.02%, and the Mean Decrease Gini model was 58.28% indicating the Mean Decrease Model was slightly better at being able to classify a record positively.

Naïve Bayes (Kevin Stewart)

Two naïve Bayes models were created to evaluate the model using different sample sizes, the Laplacian method, and the kernels method. The reduced sample size, utilizing the Laplacian method, coupled with kernels, produced the highest recall of all models in our study. The naïve Bayes' theorem uses posterior probabilities:

$$p(\mathrm{Y} = y^{*}|\mathrm{X}^{*}) = \frac{p(X^{*}|\mathrm{Y} = y^{*})p(\mathrm{Y}=y^{*})}{p(\mathrm{X}^{*})} \text{ (2)}$$

The naïve Bayes' model performance decreased by 1% from the training data set to the testing data set. First, the data set was reduced, and Laplace smoothing was applied to smooth the conditional probabilities for the various feature levels. Then, the kernel method was applied for kernel density estimation to improve performance.

The model showed that an individual is more likely to have not received the H1N1 vaccine if they have no prior knowledge of the H1N1 vaccine, use behaviors that avoid contraction, practice washing their hands, are male, are Black or Hispanic, rent as opposed to owning their home, and are not employed. In addition, the naïve Bayesian model had the best overall performance of all the models, which we measured with a sensitivity of 70.90% of true positives found.

Evaluation of models

For this study, recall/sensitivity was chosen as the success metric for determining the most appropriate model for predicting the likelihood of vaccination. This decision was made since sensitivity measure the number of true positives (vaccination), and the cost of false positives is a public health concern. Therefore, based on Table (3), Naïve Bayes 1 is selected as the most appropriate model for predicting the likelihood of vaccination with its sensitivity of 70.90%.

Table 3:

Model Evaluation for 3 models and multiple Iterations

Model	Accuracy	Specificity	Sensitivity	AUC
LogReg01	81.23%	86.60%	62.71%	83.04%
LogReg02	81.21%	86.57%	62.71%	82.04%
RandomForest	81.23%	88.54%	56.02%	82.57%
Mean Gini	80.13%	86.47%	58.28%	82.03%
Naïve Bayesian1	72.83%	74.76%	70.90%	72.83%
Naïve Bayesian2	72.51%	73.44%	69.32%	72.25%

References:

Devore, J. (2016). *Probability and Statistics for Engineering and the Sciences* (9th ed.). Boston, MA: Cengage Learning.

Hong Han, Xiaoling Guo, & Hua Yu. (2016). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference On*, 219–224. https://doi-org.sandiego.idm.oclc.org/10.1109/ICSESS.2016.7883053

Larose, C., & Larose, D. (2019). *Data Science Using Python and R*. John Wiley & Sons, Inc.

Martinez-Taboada, F., & Redondo, J. I.. (2020). *Variable importance plot (mean decrease accuracy and mean decrease Gini).*. doi: 10.1371/journal.pone.0230799.g002

Newport, F. (2021, June 4). *In U.S., 20% of Parents are unable to GET h1n1 vaccine for child*. Gallup.com. https://news.gallup.com/poll/124220/Parents-Unable-H1N1-Vaccine-Child.aspx.

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.

Appendix

Predicting H1N1 vaccine likelihood using Data Mining Methods

Luke Awino, Roberto Cancel, & Kevin Stewart

7/27/2021

Team: 6
Data set: "Flu Shot Learning:"Predict H1N1 and Seasonal Flu Vaccines"
Origin: "UCI Machine Learning Repository"
Objective: The goal is to predict the probability of individuals getting their H1N1 vaccine using behavioral and demographic information.

Data Importing and Pre-processing

Import the Training data set

Examine the structure of the data set

```
#Look at the the structure of the data
str(h1n1_df)

## 'data.frame':    26707 obs. of  36 variables:
##  $ h1n1_concern               : int  1 3 1 1 2 3 0 1 0 2 ...
##  $ h1n1_knowledge             : int  0 2 1 1 1 1 0 0 2 1 ...
##  $ behavioral_antiviral_meds  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_avoidance       : int  0 1 1 1 1 1 0 1 1 1 ...
##  $ behavioral_face_mask       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_wash_hands      : int  0 1 0 1 1 1 0 1 1 0 ...
##  $ behavioral_large_gatherings: int  0 0 0 1 1 0 0 0 1 1 ...
##  $ behavioral_outside_home    : int  1 1 0 0 0 0 0 0 1 0 ...
##  $ behavioral_touch_face      : int  1 1 0 0 1 1 0 1 1 1 ...
##  $ doctor_recc_h1n1           : int  0 0 NA 0 0 0 0 1 0 0 ...
##  $ doctor_recc_seasonal       : int  0 0 NA 1 0 1 0 0 0 0 ...
##  $ chronic_med_condition      : int  0 0 1 1 0 0 0 1 0 1 ...
##  $ child_under_6_months       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ health_worker              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ health_insurance           : int  1 1 NA NA NA NA NA 1 NA 1 ...
##  $ opinion_h1n1_vacc_effective: int  3 5 3 3 3 5 4 5 4 4 ...
##  $ opinion_h1n1_risk          : int  1 4 1 3 3 2 1 2 1 2 ...
##  $ opinion_h1n1_sick_from_vacc: int  2 4 1 5 2 1 1 1 1 2 ...
##  $ opinion_seas_vacc_effective: int  2 4 4 5 3 5 4 4 4 4 ...
##  $ opinion_seas_risk          : int  1 2 1 4 1 4 2 2 2 2 ...
##  $ opinion_seas_sick_from_vacc: int  2 4 2 1 4 4 1 1 1 2 ...
##  $ age_group                  : chr  "55 - 64 Years" "35 - 44 Years" "18 - 34 Years" "65+ Y
ears" ...
##  $ education                  : chr  "< 12 Years" "12 Years" "College Graduate" "12 Years"
...
##  $ race                       : chr  "White" "White" "White" "White" ...
##  $ sex                        : chr  "Female" "Male" "Male" "Female" ...
##  $ income_poverty             : chr  "Below Poverty" "Below Poverty" "<= $75,000, Above Pov
erty" "Below Poverty" ...
##  $ marital_status             : chr  "Not Married" "Not Married" "Not Married" "Not Married
" ...
##  $ rent_or_own                : chr  "Own" "Rent" "Own" "Rent" ...
##  $ employment_status          : chr  "Not in Labor Force" "Employed" "Employed" "Not in Lab
or Force" ...
##  $ hhs_geo_region             : chr  "oxchjgsf" "bhuqouqj" "qufhixun" "lrircsnp" ...
##  $ census_msa                 : chr  "Non-MSA" "MSA, Not Principle  City" "MSA, Not Princip
le  City" "MSA, Principle City" ...
```

```
##  $ household_adults           : int  0 0 2 0 1 2 0 2 1 0 ...
##  $ household_children         : int  0 0 0 0 0 3 0 0 0 0 ...
##  $ employment_industry        : chr  NA "pxcmvdjn" "rucpziij" NA ...
##  $ employment_occupation      : chr  NA "xgwztkwe" "xtkaffoo" NA ...
##  $ h1n1_vaccine               : int  0 0 0 0 0 0 0 1 0 0 ...
```

Examine missing values for first round of feature elimination

```
# sort missing values by count
describe(h1n1_df)
```

```
## h1n1_df
##
##  36  Variables      26707  Observations
## --------------------------------------------------------------------------------
## h1n1_concern
##        n  missing distinct     Info     Mean      Gmd
##    26615       92        4    0.901    1.618   0.9928
##
## Value          0     1     2     3
## Frequency   3296  8153 10575  4591
## Proportion 0.124 0.306 0.397 0.172
## --------------------------------------------------------------------------------
## h1n1_knowledge
##        n  missing distinct     Info     Mean      Gmd
##    26591      116        3    0.788    1.263   0.6297
##
## Value          0     1     2
## Frequency   2506 14598  9487
## Proportion 0.094 0.549 0.357
## --------------------------------------------------------------------------------
## behavioral_antiviral_meds
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26636       71        2    0.139     1301  0.04884  0.09292
##
## --------------------------------------------------------------------------------
## behavioral_avoidance
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26499      208        2    0.597    19228   0.7256   0.3982
##
## --------------------------------------------------------------------------------
## behavioral_face_mask
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26688       19        2    0.193     1841  0.06898   0.1285
##
## --------------------------------------------------------------------------------
## behavioral_wash_hands
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26665       42        2    0.432    22015   0.8256    0.288
##
## --------------------------------------------------------------------------------
## behavioral_large_gatherings
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26620       87        2     0.69     9547   0.3586   0.4601
##
## --------------------------------------------------------------------------------
## behavioral_outside_home
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26625       82        2    0.671     8981   0.3373   0.4471
##
## --------------------------------------------------------------------------------
## behavioral_touch_face
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    26579      128        2    0.656    18001   0.6773   0.4372 
## 
## --------------------------------------------------------------------------------
## doctor_recc_h1n1
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    24547     2160        2    0.515     5408   0.2203   0.3436 
## 
## --------------------------------------------------------------------------------
## doctor_recc_seasonal
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    24547     2160        2    0.663     8094   0.3297    0.442 
## 
## --------------------------------------------------------------------------------
## chronic_med_condition
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    25736      971        2    0.609     7290   0.2833   0.4061 
## 
## --------------------------------------------------------------------------------
## child_under_6_months
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    25887      820        2    0.227     2138  0.08259   0.1515 
## 
## --------------------------------------------------------------------------------
## health_worker
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    25903      804        2    0.298     2899   0.1119   0.1988 
## 
## --------------------------------------------------------------------------------
## health_insurance
##        n  missing distinct     Info      Sum     Mean      Gmd 
##    14433    12274        2    0.317    12697   0.8797   0.2116 
## 
## --------------------------------------------------------------------------------
## opinion_h1n1_vacc_effective
##        n  missing distinct     Info     Mean      Gmd 
##    26316      391        5    0.886    3.851    1.055 
## 
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
## 
## Value          1     2     3     4     5
## Frequency    886  1858  4723 11683  7166
## Proportion 0.034 0.071 0.179 0.444 0.272
## --------------------------------------------------------------------------------
## opinion_h1n1_risk
##        n  missing distinct     Info     Mean      Gmd 
##    26319      388        5    0.908    2.343    1.378 
## 
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
## 
## Value          1     2     3     4     5
## Frequency   8139  9919  1117  5394  1750
## Proportion 0.309 0.377 0.042 0.205 0.066
## --------------------------------------------------------------------------------
## opinion_h1n1_sick_from_vacc
##        n  missing distinct     Info     Mean      Gmd 
##    26312      395        5    0.907    2.358    1.455 
## 
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
## 
```

```
## Value          1     2     3     4     5
## Frequency   8998  9129   148  5850  2187
## Proportion 0.342 0.347 0.006 0.222 0.083
## --------------------------------------------------------------------------------
## opinion_seas_vacc_effective
##        n  missing distinct     Info     Mean      Gmd
##    26245      462        5    0.857    4.026    1.078
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency   1221  2206  1216 11629  9973
## Proportion 0.047 0.084 0.046 0.443 0.380
## --------------------------------------------------------------------------------
## opinion_seas_risk
##        n  missing distinct     Info     Mean      Gmd
##    26193      514        5    0.922    2.719    1.524
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency   5974  8954   677  7630  2958
## Proportion 0.228 0.342 0.026 0.291 0.113
## --------------------------------------------------------------------------------
## opinion_seas_sick_from_vacc
##        n  missing distinct     Info     Mean      Gmd
##    26170      537        5    0.875    2.118    1.374
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency  11870  7633    94  4852  1721
## Proportion 0.454 0.292 0.004 0.185 0.066
## --------------------------------------------------------------------------------
## age_group
##        n  missing distinct
##    26707        0        5
##
## lowest : 18 - 34 Years 35 - 44 Years 45 - 54 Years 55 - 64 Years 65+ Years
## highest: 18 - 34 Years 35 - 44 Years 45 - 54 Years 55 - 64 Years 65+ Years
##
## Value      18 - 34 Years 35 - 44 Years 45 - 54 Years 55 - 64 Years
## Frequency           5215          3848          5238          5563
## Proportion         0.195         0.144         0.196         0.208
##
## Value          65+ Years
## Frequency           6843
## Proportion         0.256
## --------------------------------------------------------------------------------
## education
##        n  missing distinct
##    25300     1407        4
##
## Value             < 12 Years         12 Years College Graduate     Some College
## Frequency               2363             5797            10097             7043
## Proportion             0.093            0.229            0.399            0.278
## --------------------------------------------------------------------------------
## race
##        n  missing distinct
##    26707        0        4
##
```

```
## Value                  Black          Hispanic Other or Multiple
## Frequency               2118              1755              1612
## Proportion             0.079             0.066             0.060
##
## Value                  White
## Frequency              21222
## Proportion             0.795
## --------------------------------------------------------------------------------
## sex
##        n  missing distinct
##    26707        0        2
##
## Value      Female   Male
## Frequency   15858  10849
## Proportion  0.594  0.406
## --------------------------------------------------------------------------------
## income_poverty
##        n  missing distinct
##    22284     4423        3
##
## Value      <= $75,000, Above Poverty                 > $75,000
## Frequency                      12777                      6810
## Proportion                     0.573                     0.306
##
## Value                  Below Poverty
## Frequency                       2697
## Proportion                     0.121
## --------------------------------------------------------------------------------
## marital_status
##        n  missing distinct
##    25299     1408        2
##
## Value          Married Not Married
## Frequency        13555       11744
## Proportion       0.536       0.464
## --------------------------------------------------------------------------------
## rent_or_own
##        n  missing distinct
##    24665     2042        2
##
## Value        Own  Rent
## Frequency  18736  5929
## Proportion  0.76  0.24
## --------------------------------------------------------------------------------
## employment_status
##        n  missing distinct
##    25244     1463        3
##
## Value                Employed Not in Labor Force         Unemployed
## Frequency               13560              10231               1453
## Proportion              0.537              0.405              0.058
## --------------------------------------------------------------------------------
## hhs_geo_region
##        n  missing distinct
##    26707        0       10
##
## lowest : atmpeygn bhuqouqj dqpwygqj fpwskwrf kbazzjca
## highest: lrircsnp lzgpxyit mlyzmhmf oxchjgsf qufhixun
##
## Value      atmpeygn bhuqouqj dqpwygqj fpwskwrf kbazzjca lrircsnp lzgpxyit
## Frequency      2033     2846     1126     3265     2858     2078     4297
## Proportion    0.076    0.107    0.042    0.122    0.107    0.078    0.161
##
## Value      mlyzmhmf oxchjgsf qufhixun
## Frequency      2243     2859     3102
## Proportion    0.084    0.107    0.116
## --------------------------------------------------------------------------------
## census_msa
##        n  missing distinct
##    26707        0        3
##
## Value      MSA, Not Principle  City      MSA, Principle City
## Frequency                     11645                     7864
## Proportion                    0.436                    0.294
##
## Value                        Non-MSA
## Frequency                       7198
## Proportion                     0.270
## --------------------------------------------------------------------------------
## household_adults
##        n  missing distinct     Info     Mean      Gmd
##    26458      249        4    0.807   0.8865   0.7578
##
## Value          0     1     2     3
## Frequency   8056 14474  2803  1125
## Proportion 0.304 0.547 0.106 0.043
## --------------------------------------------------------------------------------
## household_children
##        n  missing distinct     Info     Mean      Gmd
##    26458      249        4    0.645   0.5346   0.8265
##
## Value          0     1     2     3
## Frequency  18672  3175  2864  1747
## Proportion 0.706 0.120 0.108 0.066
## --------------------------------------------------------------------------------
## employment_industry
##        n  missing distinct
##    13377    13330       21
##
## lowest : arjwrbjb atmlpfrs cfqqtusy dotnnunm fcxhlnwr
## highest: vjjrobsf wlfvacwt wxleyezf xicduogh xqicxuve
## --------------------------------------------------------------------------------
## employment_occupation
##        n  missing distinct
##    13237    13470       23
##
## lowest : bxpfxfdn ccgxvspp cmhcxjea dcjcmpih dlvbwzss
## highest: vlluhbov xgwztkwe xqwwgdyp xtkaffoo xzmlyyjv
## --------------------------------------------------------------------------------
## h1n1_vaccine
##        n  missing distinct     Info      Sum     Mean      Gmd
##    26707        0        2    0.502     5674   0.2125   0.3346
##
## --------------------------------------------------------------------------------
```

Remove features with large proportion of missing data

```
#Removing employment data (since 13330/26707 or 50% of employment_industry is missing and 1347
0/26707 or 50% of employment_occupation is missing) and health_insurance (50% missing) and hhs
_geo_region to focus on Census_msa
h1n1_df <- subset(h1n1_df, select = -c(hhs_geo_region, employment_industry, employment_occupat
ion, health_insurance))
```

Review Missing Data still in df

```
# Count missing data in the data frame
sort(colSums(is.na(h1n1_df)))

##                   age_group                        race
##                           0                           0
##                         sex                  census_msa
##                           0                           0
##                h1n1_vaccine        behavioral_face_mask
##                           0                          19
##       behavioral_wash_hands   behavioral_antiviral_meds
##                          42                          71
##     behavioral_outside_home behavioral_large_gatherings
##                          82                          87
##                h1n1_concern              h1n1_knowledge
##                          92                         116
##       behavioral_touch_face        behavioral_avoidance
##                         128                         208
##            household_adults          household_children
##                         249                         249
##           opinion_h1n1_risk opinion_h1n1_vacc_effective
##                         388                         391
## opinion_h1n1_sick_from_vacc opinion_seas_vacc_effective
##                         395                         462
##           opinion_seas_risk opinion_seas_sick_from_vacc
##                         514                         537
##               health_worker        child_under_6_months
##                         804                         820
##       chronic_med_condition                   education
##                         971                        1407
##              marital_status           employment_status
##                        1408                        1463
##                 rent_or_own            doctor_recc_h1n1
##                        2042                        2160
##        doctor_recc_seasonal              income_poverty
##                        2160                        4423
```

Impute Missing Values for Categorical Variables with mode

```
h1n1_df <- h1n1_df[complete.cases(h1n1_df), ]
str(h1n1_df)

## 'data.frame':    19642 obs. of  32 variables:
##  $ h1n1_concern               : int  1 3 1 2 3 0 1 0 2 2 ...
##  $ h1n1_knowledge             : int  0 2 1 1 1 0 0 2 1 1 ...
##  $ behavioral_antiviral_meds  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_avoidance       : int  0 1 1 1 1 0 1 1 1 1 ...
##  $ behavioral_face_mask       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_wash_hands      : int  0 1 1 1 1 0 1 1 0 1 ...
##  $ behavioral_large_gatherings: int  0 0 1 1 0 0 0 1 1 1 ...
##  $ behavioral_outside_home    : int  1 1 0 0 0 0 0 1 0 0 ...
##  $ behavioral_touch_face      : int  1 1 0 1 1 0 1 1 1 0 ...
##  $ doctor_recc_h1n1           : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ doctor_recc_seasonal       : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ chronic_med_condition      : int  0 0 1 0 0 0 1 0 1 1 ...
##  $ child_under_6_months       : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ health_worker              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ opinion_h1n1_vacc_effective: int  3 5 3 3 5 4 5 4 4 4 ...
##  $ opinion_h1n1_risk          : int  1 4 3 3 2 1 2 1 2 1 ...
##  $ opinion_h1n1_sick_from_vacc: int  2 4 5 2 1 1 1 1 2 2 ...
##  $ opinion_seas_vacc_effective: int  2 4 5 3 5 4 4 4 4 5 ...
##  $ opinion_seas_risk          : int  1 2 4 1 4 2 2 2 2 4 ...
##  $ opinion_seas_sick_from_vacc: int  2 4 1 4 4 1 1 1 2 4 ...
##  $ age_group                  : chr  "55 - 64 Years" "35 - 44 Years" "65+ Years" "45 - 54 Y
ears" ...
##  $ education                  : chr  "< 12 Years" "12 Years" "12 Years" "Some College" ...
##  $ race                       : chr  "White" "White" "White" "White" ...
##  $ sex                        : chr  "Female" "Male" "Female" "Female" ...
##  $ income_poverty             : chr  "Below Poverty" "Below Poverty" "Below Poverty" "<= $7
5,000, Above Poverty" ...
##  $ marital_status             : chr  "Not Married" "Not Married" "Not Married" "Married" ..
.
##  $ rent_or_own                : chr  "Own" "Rent" "Rent" "Own" ...
##  $ employment_status          : chr  "Not in Labor Force" "Employed" "Not in Labor Force" "
Employed" ...
##  $ census_msa                 : chr  "Non-MSA" "MSA, Not Principle  City" "MSA, Principle C
ity" "MSA, Not Principle  City" ...
##  $ household_adults           : int  0 0 0 1 2 0 2 1 0 2 ...
##  $ household_children         : int  0 0 0 0 3 0 0 0 0 0 ...
##  $ h1n1_vaccine               : int  0 0 0 0 0 0 1 0 0 1 ...
```

```
#Verify that all the data is is not missing
sort(colSums(is.na(h1n1_df)))
```

```
##                h1n1_concern              h1n1_knowledge
##                           0                           0
##   behavioral_antiviral_meds        behavioral_avoidance
##                           0                           0
##        behavioral_face_mask       behavioral_wash_hands
##                           0                           0
## behavioral_large_gatherings     behavioral_outside_home
##                           0                           0
##       behavioral_touch_face            doctor_recc_h1n1
##                           0                           0
##        doctor_recc_seasonal       chronic_med_condition
##                           0                           0
##        child_under_6_months               health_worker
##                           0                           0
## opinion_h1n1_vacc_effective           opinion_h1n1_risk
##                           0                           0
## opinion_h1n1_sick_from_vacc opinion_seas_vacc_effective
##                           0                           0
##           opinion_seas_risk opinion_seas_sick_from_vacc
##                           0                           0
##                   age_group                   education
##                           0                           0
##                        race                         sex
##                           0                           0
##              income_poverty              marital_status
##                           0                           0
##                 rent_or_own           employment_status
##                           0                           0
##                  census_msa            household_adults
##                           0                           0
##          household_children                h1n1_vaccine
##                           0                           0
```

Transform the features

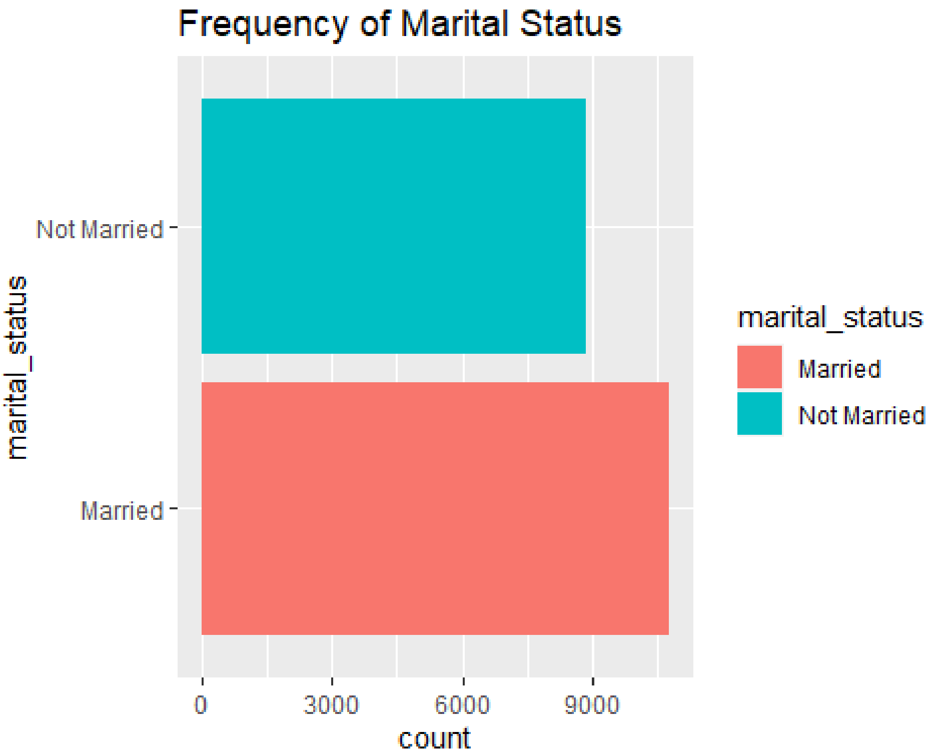
```
#converting categorical variables to factors
h1n1_df$education <- as.factor(h1n1_df$education)
```

```
h1n1_df$race <- as.factor(h1n1_df$race)
h1n1_df$sex <- as.factor(h1n1_df$sex)
h1n1_df$age_group <- as.factor(h1n1_df$age_group)
h1n1_df$income_poverty <- as.factor(h1n1_df$income_poverty)
h1n1_df$marital_status <- as.factor(h1n1_df$marital_status)
h1n1_df$rent_or_own <- as.factor(h1n1_df$rent_or_own)
h1n1_df$employment_status <- as.factor(h1n1_df$employment_status)
#converting integers discrete variables to factors
h1n1_df$h1n1_concern <- as.factor(h1n1_df$h1n1_concern)
h1n1_df$h1n1_knowledge <- as.factor(h1n1_df$h1n1_knowledge)
h1n1_df$behavioral_antiviral_meds <- as.factor(h1n1_df$behavioral_antiviral_meds)
h1n1_df$behavioral_avoidance <- as.factor(h1n1_df$behavioral_avoidance)
h1n1_df$behavioral_face_mask <- as.factor(h1n1_df$behavioral_face_mask)
h1n1_df$behavioral_wash_hands <- as.factor(h1n1_df$behavioral_wash_hands)
h1n1_df$behavioral_large_gatherings <- as.factor(h1n1_df$behavioral_large_gatherings)
h1n1_df$behavioral_outside_home <- as.factor(h1n1_df$behavioral_outside_home)
h1n1_df$behavioral_outside_home <- as.factor(h1n1_df$behavioral_touch_face)
h1n1_df$behavioral_touch_face <- as.factor(h1n1_df$behavioral_touch_face)
h1n1_df$doctor_recc_h1n1 <- as.factor(h1n1_df$doctor_recc_h1n1)
h1n1_df$doctor_recc_seasonal <- as.factor(h1n1_df$doctor_recc_seasonal)
h1n1_df$chronic_med_condition <- as.factor(h1n1_df$chronic_med_condition)
h1n1_df$child_under_6_months <- as.factor(h1n1_df$child_under_6_months)
h1n1_df$health_worker <- as.factor(h1n1_df$health_worker)
h1n1_df$opinion_h1n1_vacc_effective <- as.factor(h1n1_df$opinion_h1n1_vacc_effective)
h1n1_df$opinion_h1n1_risk <- as.factor(h1n1_df$opinion_h1n1_risk)
h1n1_df$opinion_h1n1_sick_from_vacc <- as.factor(h1n1_df$opinion_h1n1_sick_from_vacc)
h1n1_df$opinion_seas_vacc_effective <- as.factor(h1n1_df$opinion_seas_vacc_effective)
h1n1_df$opinion_seas_risk <- as.factor(h1n1_df$opinion_seas_risk)
h1n1_df$opinion_seas_sick_from_vacc <- as.factor(h1n1_df$opinion_seas_sick_from_vacc)
h1n1_df$household_adults <- as.factor(h1n1_df$household_adults)
h1n1_df$household_children <- as.factor(h1n1_df$household_children)
h1n1_df$census_msa <- as.factor(h1n1_df$census_msa)

clean_data <- h1n1_df
str(clean_data)

## 'data.frame':    19642 obs. of  32 variables:
##  $ h1n1_concern               : Factor w/ 4 levels "0","1","2","3": 2 4 2 3 4 1 2 1 3 3 ...
##  $ h1n1_knowledge             : Factor w/ 3 levels "0","1","2": 1 3 2 2 2 1 1 3 2 2 ...
##  $ behavioral_antiviral_meds  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_avoidance       : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 2 2 ...
##  $ behavioral_face_mask       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_wash_hands      : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 1 2 ...
##  $ behavioral_large_gatherings: Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 2 2 2 ...
##  $ behavioral_outside_home    : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 1 ...
##  $ behavioral_touch_face      : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 1 ...
##  $ doctor_recc_h1n1           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ doctor_recc_seasonal       : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 1 1 1 ...
##  $ chronic_med_condition      : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 2 2 ...
##  $ child_under_6_months       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ health_worker              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ opinion_h1n1_vacc_effective: Factor w/ 5 levels "1","2","3","4",..: 3 5 3 3 5 4 5 4 4 4
...
##  $ opinion_h1n1_risk          : Factor w/ 5 levels "1","2","3","4",..: 1 4 3 3 2 1 2 1 2 1
...
##  $ opinion_h1n1_sick_from_vacc: Factor w/ 5 levels "1","2","3","4",..: 2 4 5 2 1 1 1 1 2 2
...
##  $ opinion_seas_vacc_effective: Factor w/ 5 levels "1","2","3","4",..: 2 4 5 3 5 4 4 4 4 5
...
##  $ opinion_seas_risk          : Factor w/ 5 levels "1","2","3","4",..: 1 2 4 1 4 2 2 2 2 4
...
```

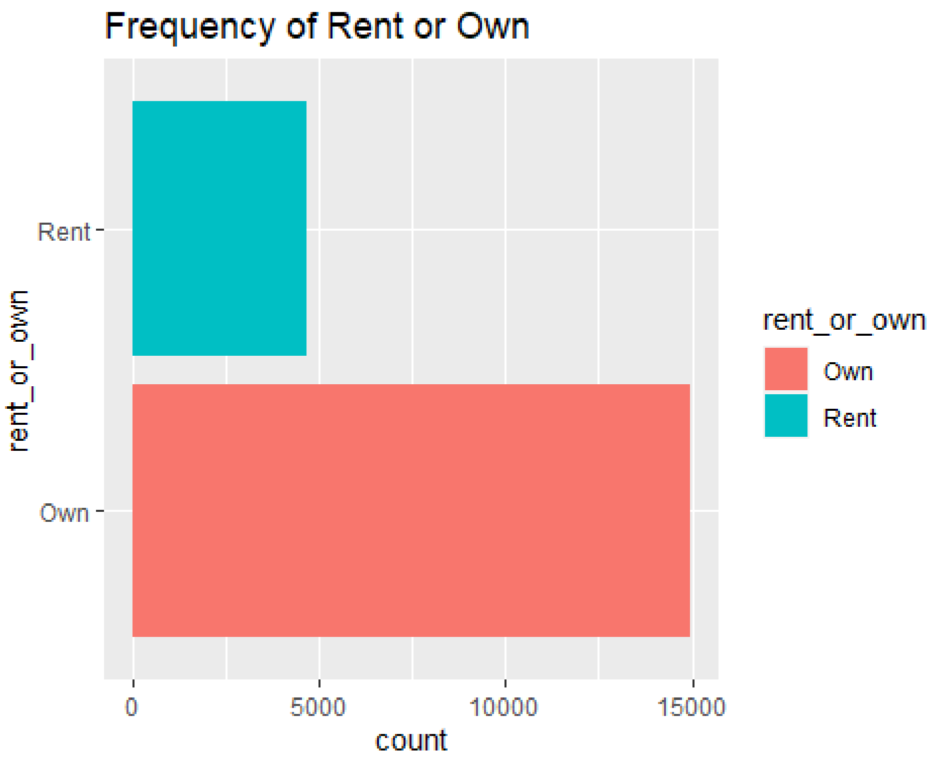
```
##  $ opinion_seas_sick_from_vacc: Factor w/ 5 levels "1","2","3","4",..: 2 4 1 4 4 1 1 1 2 4
...
##  $ age_group                  : Factor w/ 5 levels "18 - 34 Years",..: 4 2 5 3 5 4 3 3 4 3
...
##  $ education                  : Factor w/ 4 levels "< 12 Years","12 Years",..: 1 2 2 4 2 1
4 3 2 2 ...
##  $ race                       : Factor w/ 4 levels "Black","Hispanic",..: 4 4 4 4 4 4 4 4 4
4 ...
##  $ sex                        : Factor w/ 2 levels "Female","Male": 1 2 1 1 2 2 1 2 2 2 ...
##  $ income_poverty             : Factor w/ 3 levels "<= $75,000, Above Poverty",..: 3 3 3 1
1 1 1 2 1 1 ...
##  $ marital_status             : Factor w/ 2 levels "Married","Not Married": 2 2 2 1 1 2 1 1
2 1 ...
##  $ rent_or_own                : Factor w/ 2 levels "Own","Rent": 1 2 2 1 1 1 1 1 1 2 ...
##  $ employment_status          : Factor w/ 3 levels "Employed","Not in Labor Force",..: 2 1
2 1 1 1 1 1 2 1 ...
##  $ census_msa                 : Factor w/ 3 levels "MSA, Not Principle  City",..: 3 1 2 1 2
1 3 1 1 1 ...
##  $ household_adults           : Factor w/ 4 levels "0","1","2","3": 1 1 1 2 3 1 3 2 1 3 ...
##  $ household_children         : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 4 1 1 1 1 1 ...
##  $ h1n1_vaccine               : int  0 0 0 0 0 0 1 0 0 1 ...
```

#Visualize categorical variables

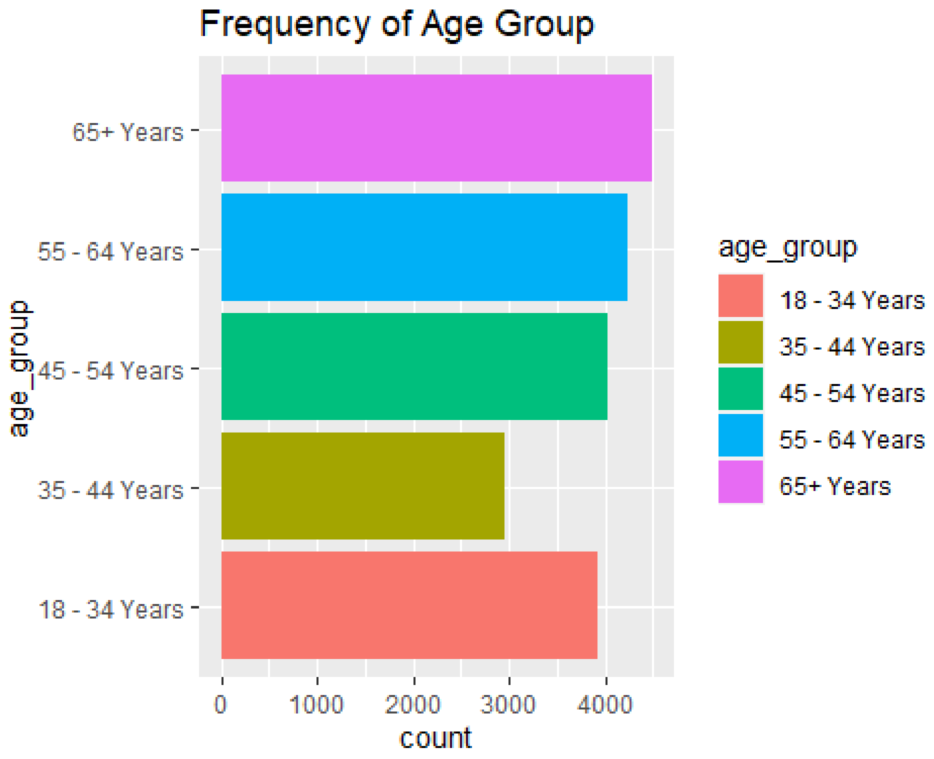
```
#graph
ggplot(clean_data, aes(marital_status)) + geom_bar(aes(fill = marital_status)) + coord_flip()
+ ggtitle("Frequency of Marital Status")
```



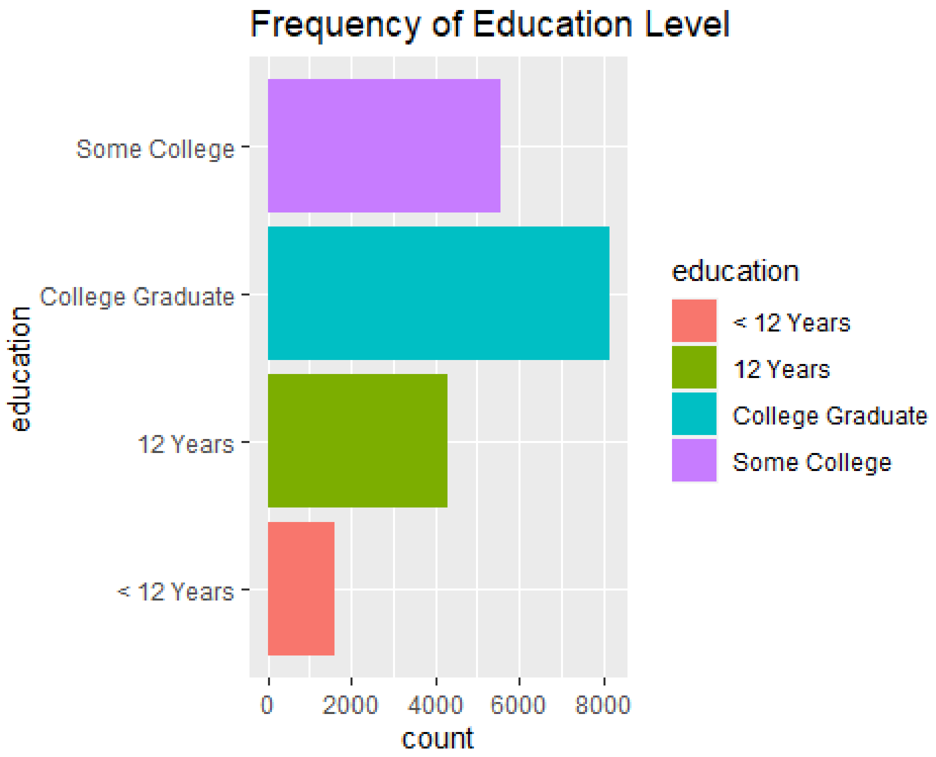
```
ggplot(clean_data, aes(rent_or_own)) + geom_bar(aes(fill = rent_or_own)) + coord_flip()+ ggtit
le("Frequency of Rent or Own")
```



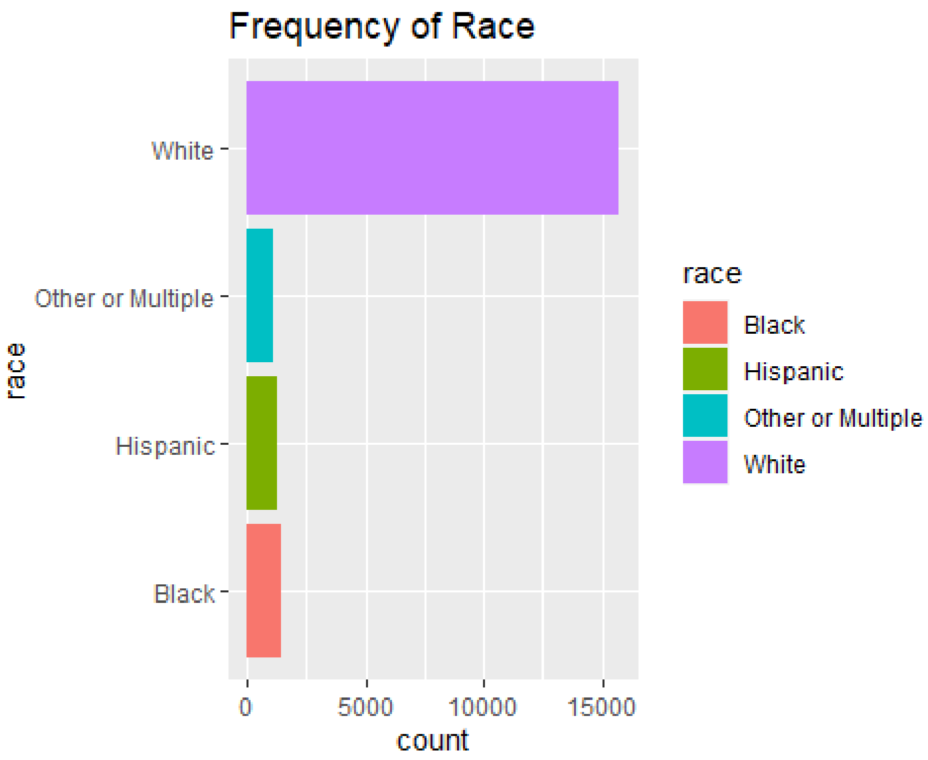
```
ggplot(clean_data, aes(age_group)) + geom_bar(aes(fill = age_group)) + coord_flip() + ggtitle(
"Frequency of Age Group")
```



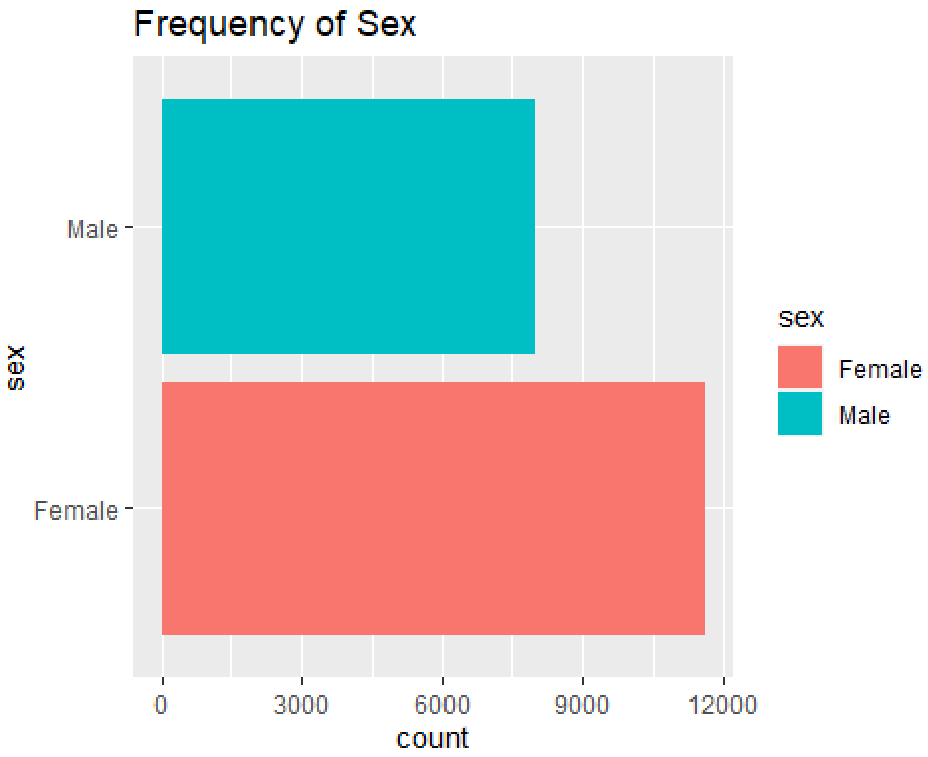
```
ggplot(clean_data, aes(education)) + geom_bar(aes(fill = education)) + coord_flip()+ ggtitle("
Frequency of Education Level")
```



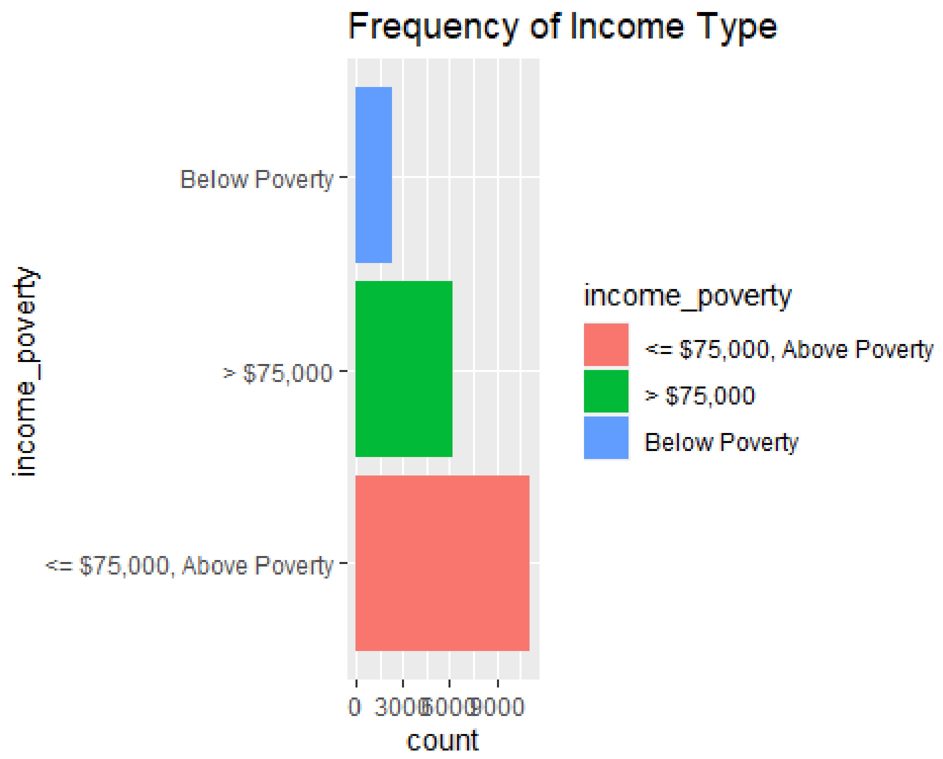
```
ggplot(clean_data, aes(race)) + geom_bar(aes(fill = race)) + coord_flip()+ ggtitle("Frequency
of Race")
```



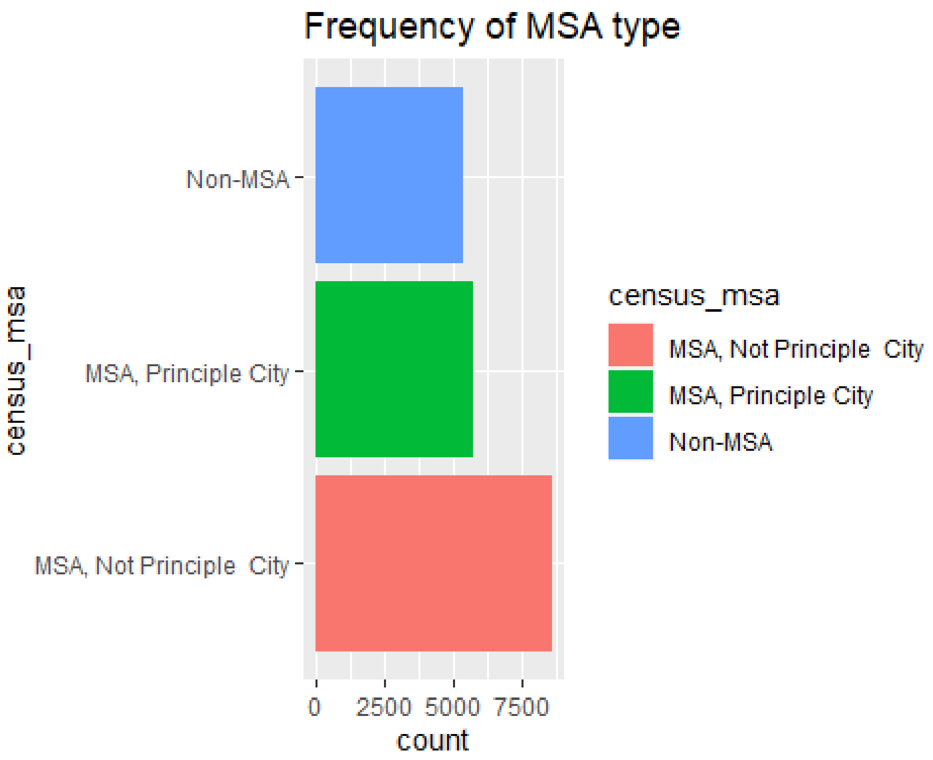
```
ggplot(clean_data, aes(sex)) + geom_bar(aes(fill = sex)) + coord_flip()+ ggtitle("Frequency of
Sex")
```



```
ggplot(clean_data, aes(income_poverty)) + geom_bar(aes(fill = income_poverty)) + coord_flip()
+ ggtitle("Frequency of Income Type")
```

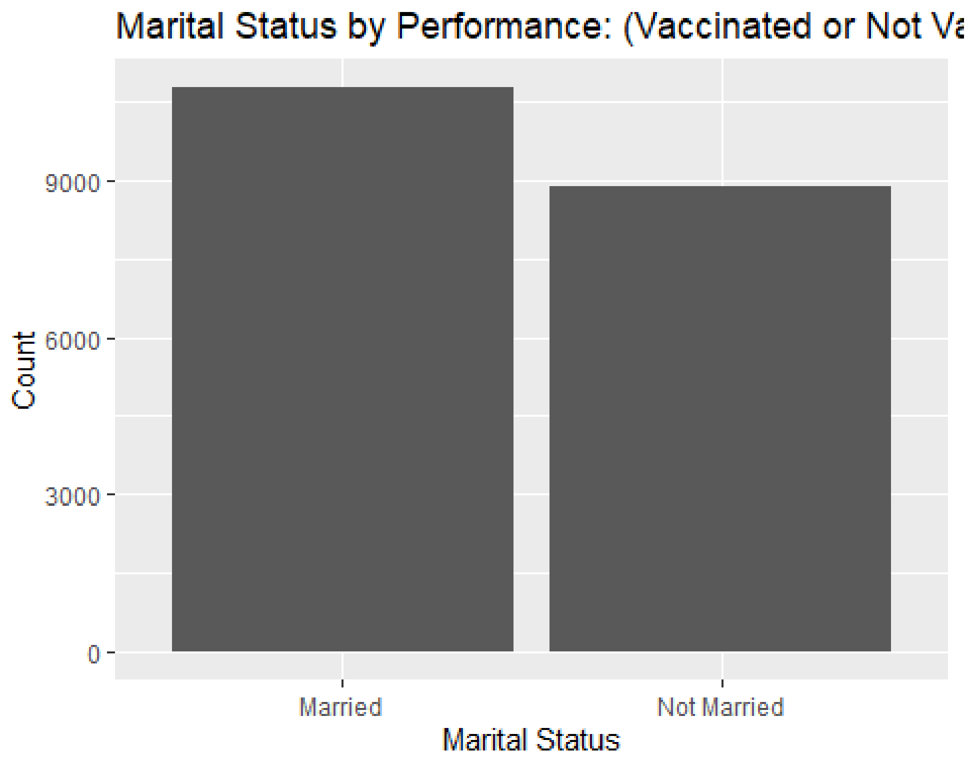


```
ggplot(clean_data, aes(census_msa)) + geom_bar(aes(fill = census_msa)) + coord_flip() + ggtitl
e("Frequency of MSA type")
```



# Visualize the categorical variables as functions of h1n1_vaccine

```
library(forcats)
ggplot(clean_data, aes(fct_infreq(marital_status))) +
geom_bar(stat="count", aes(fill= h1n1_vaccine)) +
labs(x = "Marital Status", y = "Count") +
ggtitle("Marital Status by Performance: (Vaccinated or Not Vaccinated)")
```



FEATURE ENGINEERING *Re-express Categorical Variables and convert to numerical*

```
unique(clean_data$marital_status)

## [1] Not Married Married
## Levels: Married Not Married

unique(clean_data$income_poverty)

## [1] Below Poverty             <= $75,000, Above Poverty
## [3] > $75,000
## Levels: <= $75,000, Above Poverty > $75,000 Below Poverty

unique(clean_data$rent_or_own)

## [1] Own  Rent
## Levels: Own Rent

unique(clean_data$education)

## [1] < 12 Years       12 Years         Some College     College Graduate
## Levels: < 12 Years 12 Years College Graduate Some College

unique(clean_data$employment_status)

## [1] Not in Labor Force Employed           Unemployed
## Levels: Employed Not in Labor Force Unemployed

# Re-expressing categorical variables as a value
marital_num <- revalue(x = clean_data$marital_status, replace = c("Not Married" = 0, "Married"
= 1))
clean_data$marital_numeric <- as.numeric(levels(marital_num))[marital_num]
```

```
# Re-expressing census_msa
census_msa <- as.factor(clean_data)
census_msa_num <- census_msa_num <- revalue(x = clean_data$census_msa, replace = c("Non-MSA" =
0, "MSA, Not Principle  City" = 1, "MSA, Principle City" = 2))
clean_data$census_msa_numeric <- as.numeric(levels(census_msa_num))[census_msa_num]

# Re-expressing age as numeric
unique(clean_data$age_group)

## [1] 55 - 64 Years 35 - 44 Years 65+ Years     45 - 54 Years 18 - 34 Years
## 5 Levels: 18 - 34 Years 35 - 44 Years 45 - 54 Years ... 65+ Years

length(unique(clean_data$age_group))

## [1] 5

age_num <- revalue(x = clean_data$age_group, replace = c("18 - 34 Years" = 0, "35 - 44 Years"
= 1, "45 - 54 Years" = 2, "55 - 64 Years" = 3, "65+ Years" = 4))
# convert age_num to numeric
clean_data$age_numeric <- as.numeric(levels(age_num))[age_num]

#Re-express sex as numeric
sex_num <- revalue(x = clean_data$sex, replace = c("Female" = 0, "Male" = 1))
clean_data$sex_numeric <- as.numeric(levels(sex_num))[sex_num]

#convert race to numeric
unique(clean_data$race)

## [1] White             Black             Hispanic          Other or Multiple
## Levels: Black Hispanic Other or Multiple White

race_num <- revalue(x = clean_data$race, replace = c("White" = 0, "Black" = 1, "Other or Multi
ple" = 2, "Hispanic" = 3))
clean_data$race_numeric <- as.numeric(levels(race_num))[race_num]

#converting income_poverty to numeric
income_poverty_num <- revalue(x = clean_data$income_poverty, replace = c("Below Poverty" = 0,
"<= $75,000, Above Poverty" = 1, "> $75,000" = 2))
clean_data$income_poverty_numeric <- as.numeric(levels(income_poverty_num))[income_poverty_num
]

#Re-expressing categorical variables
unique(clean_data$rent_or_own)

## [1] Own  Rent
## Levels: Own Rent

rent_or_own_num <- revalue(x = clean_data$rent_or_own, replace = c("Own" = 0, "Rent" = 1))
clean_data$rent_or_own_numeric <- as.numeric(levels(rent_or_own_num))[rent_or_own_num]

#Re-expressing categorical variables
unique(clean_data$education)

## [1] < 12 Years       12 Years         Some College     College Graduate
## Levels: < 12 Years 12 Years College Graduate Some College

education_num <- revalue(x = clean_data$education, replace = c("< 12 Years" = 0, "12 Years"= 1
, "College Graduate" = 2, "Some College"= 3))
clean_data$education_numeric <- as.numeric(levels(education_num)) [education_num]
```

```
#Re-expressing categorical variables
unique(clean_data$employment_status)

## [1] Not in Labor Force Employed           Unemployed
## Levels: Employed Not in Labor Force Unemployed

employment_num <- revalue(x = clean_data$employment_status, replace = c("Unemployed" = 0, "Not
in Labor Force" = 1, "Employed" = 2))
clean_data$employment_numeric <- as.numeric(levels(employment_num)) [employment_num]



#Re-express categorical variables
str(clean_data)

## 'data.frame':    19642 obs. of  41 variables:
##  $ h1n1_concern              : Factor w/ 4 levels "0","1","2","3": 2 4 2 3 4 1 2 1 3 3 ...
##  $ h1n1_knowledge            : Factor w/ 3 levels "0","1","2": 1 3 2 2 2 1 1 3 2 2 ...
##  $ behavioral_antiviral_meds  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_avoidance      : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 2 2 ...
##  $ behavioral_face_mask      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_wash_hands     : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 1 2 ...
##  $ behavioral_large_gatherings: Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 2 2 2 ...
##  $ behavioral_outside_home   : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 1 ...
##  $ behavioral_touch_face     : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 1 ...
##  $ doctor_recc_h1n1          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ doctor_recc_seasonal      : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 1 1 1 ...
##  $ chronic_med_condition     : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 2 2 ...
##  $ child_under_6_months      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ health_worker             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ opinion_h1n1_vacc_effective: Factor w/ 5 levels "1","2","3","4",..: 3 5 3 3 5 4 5 4 4 4
...
##  $ opinion_h1n1_risk         : Factor w/ 5 levels "1","2","3","4",..: 1 4 3 3 2 1 2 1 2 1
...
##  $ opinion_h1n1_sick_from_vacc: Factor w/ 5 levels "1","2","3","4",..: 2 4 5 2 1 1 1 1 2 2
...
##  $ opinion_seas_vacc_effective: Factor w/ 5 levels "1","2","3","4",..: 2 4 5 3 5 4 4 4 4 5
...
##  $ opinion_seas_risk         : Factor w/ 5 levels "1","2","3","4",..: 1 2 4 1 4 2 2 2 2 4
...
##  $ opinion_seas_sick_from_vacc: Factor w/ 5 levels "1","2","3","4",..: 2 4 1 4 4 1 1 1 2 4
...
##  $ age_group                 : Factor w/ 5 levels "18 - 34 Years",..: 4 2 5 3 5 4 3 3 4 3
...
##  $ education                 : Factor w/ 4 levels "< 12 Years","12 Years",..: 1 2 2 4 2 1
4 3 2 2 ...
##  $ race                      : Factor w/ 4 levels "Black","Hispanic",..: 4 4 4 4 4 4 4 4 4
4 ...
##  $ sex                       : Factor w/ 2 levels "Female","Male": 1 2 1 1 2 2 1 2 2 2 ...
##  $ income_poverty            : Factor w/ 3 levels "<= $75,000, Above Poverty",..: 3 3 3 1
1 1 1 2 1 1 ...
##  $ marital_status            : Factor w/ 2 levels "Married","Not Married": 2 2 2 1 1 2 1 1
2 1 ...
##  $ rent_or_own               : Factor w/ 2 levels "Own","Rent": 1 2 2 1 1 1 1 1 1 2 ...
##  $ employment_status         : Factor w/ 3 levels "Employed","Not in Labor Force",..: 2 1
2 1 1 1 1 1 2 1 ...
##  $ census_msa                : Factor w/ 3 levels "MSA, Not Principle  City",..: 3 1 2 1 2
1 3 1 1 1 ...
##  $ household_adults          : Factor w/ 4 levels "0","1","2","3": 1 1 1 2 3 1 3 2 1 3 ...
##  $ household_children        : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 4 1 1 1 1 1 ...
##  $ h1n1_vaccine              : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ marital_numeric           : num  0 0 0 1 1 0 1 1 0 1 ...
##  $ census_msa_numeric         : num  0 1 2 1 2 1 0 1 1 1 ...
##  $ age_numeric                : num  3 1 4 2 4 3 2 2 3 2 ...
##  $ sex_numeric                : num  0 1 0 0 1 1 0 1 1 1 ...
##  $ race_numeric               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ income_poverty_numeric     : num  0 0 0 1 1 1 1 2 1 1 ...
##  $ rent_or_own_numeric        : num  0 1 1 0 0 0 0 0 0 1 ...
##  $ education_numeric          : num  0 1 1 3 1 0 3 2 1 1 ...
##  $ employment_numeric         : num  1 2 1 2 2 2 2 2 1 2 ...
```

Drop the Categorical Variables and Seasonal Flu data since we're focusing on H1N1 vaccines

```
clean_data1 <- subset(clean_data, select = -c( age_group, education, race, sex, income_poverty
, marital_status, rent_or_own, employment_status, census_msa, doctor_recc_seasonal, opinion_se
as_vacc_effective, opinion_seas_risk, opinion_seas_sick_from_vacc))
```

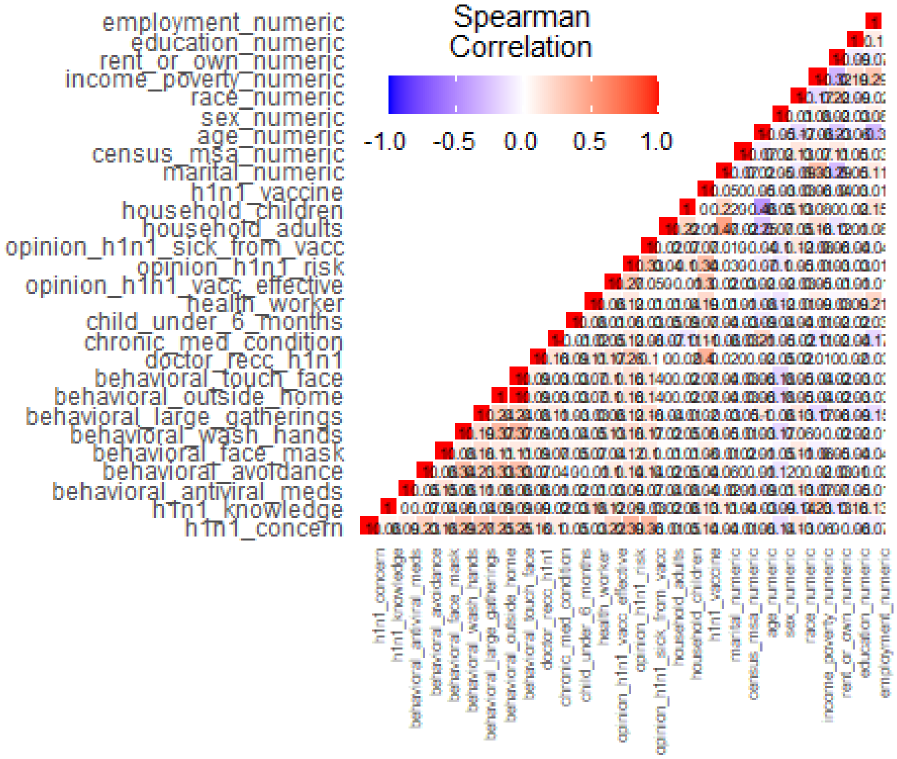
Convert all variables to numeric after transformations

```
prep_data <- mutate_all(clean_data1, function(clean_data)as.numeric(clean_data))
str(prep_data)

## 'data.frame':    19642 obs. of  28 variables:
##  $ h1n1_concern               : num  2 4 2 3 4 1 2 1 3 3 ...
##  $ h1n1_knowledge             : num  1 3 2 2 2 1 1 3 2 2 ...
##  $ behavioral_antiviral_meds  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_avoidance       : num  1 2 2 2 2 1 2 2 2 2 ...
##  $ behavioral_face_mask       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_wash_hands      : num  1 2 2 2 2 1 2 2 1 2 ...
##  $ behavioral_large_gatherings: num  1 1 2 2 1 1 1 2 2 2 ...
##  $ behavioral_outside_home    : num  2 2 1 2 2 1 2 2 2 1 ...
##  $ behavioral_touch_face      : num  2 2 1 2 2 1 2 2 2 1 ...
##  $ doctor_recc_h1n1           : num  1 1 1 1 1 1 2 1 1 1 ...
##  $ chronic_med_condition      : num  1 1 2 1 1 1 2 1 2 2 ...
##  $ child_under_6_months       : num  1 1 1 1 1 1 1 1 1 2 ...
##  $ health_worker              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ opinion_h1n1_vacc_effective: num  3 5 3 3 5 4 5 4 4 4 ...
##  $ opinion_h1n1_risk          : num  1 4 3 3 2 1 2 1 2 1 ...
##  $ opinion_h1n1_sick_from_vacc: num  2 4 5 2 1 1 1 1 2 2 ...
##  $ household_adults           : num  1 1 1 2 3 1 3 2 1 3 ...
##  $ household_children         : num  1 1 1 1 4 1 1 1 1 1 ...
##  $ h1n1_vaccine               : num  0 0 0 0 0 0 1 0 0 1 ...
##  $ marital_numeric            : num  0 0 0 1 1 0 1 1 0 1 ...
##  $ census_msa_numeric         : num  0 1 2 1 2 1 0 1 1 1 ...
##  $ age_numeric                : num  3 1 4 2 4 3 2 2 3 2 ...
##  $ sex_numeric                : num  0 1 0 0 1 1 0 1 1 1 ...
##  $ race_numeric               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ income_poverty_numeric     : num  0 0 0 1 1 1 1 2 1 1 ...
##  $ rent_or_own_numeric        : num  0 1 1 0 0 0 0 0 0 1 ...
##  $ education_numeric          : num  0 1 1 3 1 0 3 2 1 1 ...
##  $ employment_numeric         : num  1 2 1 2 2 2 2 2 1 2 ...

#View updated prep_data
str(prep_data)

## 'data.frame':    19642 obs. of  28 variables:
##  $ h1n1_concern               : num  2 4 2 3 4 1 2 1 3 3 ...
##  $ h1n1_knowledge             : num  1 3 2 2 2 1 1 3 2 2 ...
##  $ behavioral_antiviral_meds  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_avoidance       : num  1 2 2 2 2 1 2 2 2 2 ...
##  $ behavioral_face_mask       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ behavioral_wash_hands      : num  1 2 2 2 2 1 2 2 1 2 ...
##  $ behavioral_large_gatherings: num  1 1 2 2 1 1 1 2 2 2 ...
##  $ behavioral_outside_home    : num  2 2 1 2 2 1 2 2 2 1 ...
##  $ behavioral_touch_face      : num  2 2 1 2 2 1 2 2 2 1 ...
##  $ doctor_recc_h1n1           : num  1 1 1 1 1 1 2 1 1 1 ...
##  $ chronic_med_condition      : num  1 1 2 1 1 1 2 1 2 2 ...
##  $ child_under_6_months       : num  1 1 1 1 1 1 1 1 1 2 ...
##  $ health_worker              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ opinion_h1n1_vacc_effective: num  3 5 3 3 5 4 5 4 4 4 ...
##  $ opinion_h1n1_risk          : num  1 4 3 3 2 1 2 1 2 1 ...
##  $ opinion_h1n1_sick_from_vacc: num  2 4 5 2 1 1 1 1 2 2 ...
##  $ household_adults           : num  1 1 1 2 3 1 3 2 1 3 ...
##  $ household_children         : num  1 1 1 1 4 1 1 1 1 1 ...
##  $ h1n1_vaccine               : num  0 0 0 0 0 0 1 0 0 1 ...
##  $ marital_numeric            : num  0 0 0 1 1 0 1 1 0 1 ...
##  $ census_msa_numeric         : num  0 1 2 1 2 1 0 1 1 1 ...
##  $ age_numeric                : num  3 1 4 2 4 3 2 2 3 2 ...
##  $ sex_numeric                : num  0 1 0 0 1 1 0 1 1 1 ...
##  $ race_numeric               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ income_poverty_numeric     : num  0 0 0 1 1 1 1 2 1 1 ...
##  $ rent_or_own_numeric        : num  0 1 1 0 0 0 0 0 0 1 ...
##  $ education_numeric          : num  0 1 1 3 1 0 3 2 1 1 ...
##  $ employment_numeric         : num  1 2 1 2 2 2 2 2 1 2 ...
```

#Final preparation before modeling

Check for multicollinearity in features

```
# calculate correlation matrix
cormat <- round(cor(prep_data, method = "spearman"), 2)
# Melt the cormat
melted_cormat <- melt(cormat)
# Get lower triangle of the correlation matrix
  get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat)] <- NA
    return(cormat)
  }
  # Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }
#Get upper tri
upper_tri <- get_upper_tri(cormat)

#Create clearer correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)

#Create heat map
ggheatmap <- ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Spearman\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 90, vjust = 1,
    size =6, hjust = 1))+
 coord_fixed()

#Add Coefficients
ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 2) +
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.position = c(0.6, 0.7),
  legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                title.position = "top", title.hjust = 0.5))
```



Scale all the features

```
#Scale ordinal features
max = apply(prep_data,2, max)
min = apply(prep_data,2, min)
prep_data = as.data.frame(scale(prep_data, center = min, scale = max - min))
```

Partition the data

```
#Set Seed and determine dimensions of data set
set.seed(654)
n <- dim(prep_data)
n

## [1] 19642    28

# Split the data into 75% train and 25% test
dt = sort(sample(nrow(prep_data), nrow(prep_data)*.75))

prep_train <- prep_data[dt,]
prep_test <- prep_data[-dt,]
```

Balance the training set

```
#Count number of records in train set
dim(prep_train)

## [1] 14731    28

#Count number of records in test set
dim(prep_test)

## [1] 4911   28
```

Identify number of h1n1_vaccine is True in training set

```
length(which(prep_train$h1n1_vaccine == "1"))

## [1] 3409
```

There are 20030 records in the training data set, of which 4273 have a h1n1_vaccine of True/1 - this means only 21% of the training set has h1n1_vaccine of True/1. We would like to balance the training set to a 50/50 of h1n1_vaccine True/1 and False/0.

To reach this x = ((.5*14731)-3409)/.5 x = 7913

We therefore need to over sample the h1n1_vaccine True/1 records by 11544 to balance our data set.

Balance the training data set for the imbalance in H1N1_vaccine

```
# Define the records to be sample from
to.resample <- which(prep_train$h1n1_vaccine == "1")
# Build a sample of size 11,544 from identified records
our.resample <- sample(x = to.resample, size = 7913, replace = TRUE)
our.resample <- prep_train[our.resample,]
# Bind re-sampled records with training data
prep_train_rebal <- rbind(prep_train, our.resample)
# Build Table of Response Counts and Proportions
t.v1 <- table(prep_train_rebal$h1n1_vaccine)
t.v2 <- rbind(t.v1, round(prop.table(t.v1), 2))
colnames(t.v2) <- c("h1n1_vaccine = False/0", "h1n1_vaccine = True/1")
rownames(t.v2) <- c("Count", "Proportion")
t.v2

##            h1n1_vaccine = False/0 h1n1_vaccine = True/1
## Count                     11322.0               11322.0
## Proportion                    0.5                   0.5

str(prep_train_rebal)

## 'data.frame':    22644 obs. of  28 variables:
##  $ h1n1_concern               : num  0.333 1 0.333 0.667 1 ...
##  $ h1n1_knowledge             : num  0 1 0.5 0.5 0.5 0 1 0.5 0.5 1 ...
##  $ behavioral_antiviral_meds  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_avoidance       : num  0 1 1 1 1 0 1 1 1 1 ...
##  $ behavioral_face_mask       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavioral_wash_hands      : num  0 1 1 1 1 0 1 0 1 1 ...
##  $ behavioral_large_gatherings: num  0 0 1 1 0 0 1 1 1 0 ...
##  $ behavioral_outside_home    : num  1 1 0 1 1 0 1 1 0 0 ...
##  $ behavioral_touch_face      : num  1 1 0 1 1 0 1 1 0 0 ...
##  $ doctor_recc_h1n1           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ chronic_med_condition      : num  0 0 1 0 0 0 0 1 1 0 ...
##  $ child_under_6_months       : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ health_worker              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ opinion_h1n1_vacc_effective: num  0.5 1 0.5 0.5 1 0.75 0.75 0.75 0.75 0.5 ...
##  $ opinion_h1n1_risk          : num  0 0.75 0.5 0.5 0.25 0 0 0.25 0 0.25 ...
##  $ opinion_h1n1_sick_from_vacc: num  0.25 0.75 1 0.25 0 0 0 0.25 0.25 0.25 ...
##  $ household_adults           : num  0 0 0 0.333 0.667 ...
##  $ household_children         : num  0 0 0 0 1 ...
##  $ h1n1_vaccine               : num  0 0 0 0 0 0 0 0 1 1 ...
##  $ marital_numeric            : num  0 0 0 1 1 0 1 0 1 1 ...
##  $ census_msa_numeric         : num  0 0.5 1 0.5 1 0.5 0.5 0.5 0.5 0 ...
##  $ age_numeric                : num  0.75 0.25 1 0.5 1 0.75 0.5 0.75 0.5 0.75 ...
##  $ sex_numeric                : num  0 1 0 0 1 1 1 1 1 1 ...
##  $ race_numeric               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ income_poverty_numeric     : num  0 0 0 0.5 0.5 0.5 1 0.5 0.5 1 ...
##  $ rent_or_own_numeric        : num  0 1 1 0 0 0 0 0 1 0 ...
##  $ education_numeric          : num  0 0.333 0.333 1 0.333 ...
##  $ employment_numeric         : num  0.5 1 0.5 1 1 1 1 0.5 1 1 ...
```

#Modeling# *Logistic Regression*

```
#Build and train baseline model with all remaining features

logreg01 <- glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_
meds + behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands + behavioral_large_
gatherings + behavioral_outside_home + doctor_recc_h1n1 + chronic_med_condition + child_under_
6_months + health_worker + opinion_h1n1_vacc_effective + opinion_h1n1_risk + opinion_h1n1_sick
_from_vacc + household_adults + household_children + marital_numeric + census_msa_numeric + ag
e_numeric + sex_numeric + race_numeric + income_poverty_numeric + rent_or_own_numeric + educat
ion_numeric + employment_numeric,
    data = prep_train_rebal, family = binomial(link = "logit"))

summary(logreg01)

##
## Call:
## glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge +
##     behavioral_antiviral_meds + behavioral_avoidance + behavioral_face_mask +
##     behavioral_wash_hands + behavioral_large_gatherings + behavioral_outside_home +
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months +
##     health_worker + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + household_adults + household_children +
##     marital_numeric + census_msa_numeric + age_numeric + sex_numeric +
##     race_numeric + income_poverty_numeric + rent_or_own_numeric +
##     education_numeric + employment_numeric, family = binomial(link = "logit"),
##     data = prep_train_rebal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.87478  -0.79266  -0.00305   0.78665   2.87062
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -4.08998    0.11984 -34.130  < 2e-16 ***
## h1n1_concern                  -0.26952    0.06635  -4.062 4.86e-05 ***
## h1n1_knowledge                 0.34285    0.05889   5.822 5.82e-09 ***
## behavioral_antiviral_meds      0.13732    0.07469   1.839 0.065986 .
## behavioral_avoidance          -0.08552    0.04180  -2.046 0.040758 *
## behavioral_face_mask           0.14651    0.06571   2.230 0.025781 *
## behavioral_wash_hands        0.10647    0.05173   2.058 0.039597 *  
## behavioral_large_gatherings -0.24340    0.03795  -6.414 1.42e-10 ***
## behavioral_outside_home     -0.04235    0.03981  -1.064 0.287455    
## doctor_recc_h1n1             1.70075    0.03759  45.247  < 2e-16 ***
## chronic_med_condition        0.14094    0.03731   3.777 0.000158 ***
## child_under_6_months         0.23488    0.05871   4.001 6.31e-05 ***
## health_worker                1.01000    0.05039  20.044  < 2e-16 ***
## opinion_h1n1_vacc_effective  2.62019    0.08218  31.885  < 2e-16 ***
## opinion_h1n1_risk            1.88648    0.05679  33.218  < 2e-16 ***
## opinion_h1n1_sick_from_vacc -0.02167    0.05353  -0.405 0.685618    
## household_adults            -0.08949    0.07521  -1.190 0.234103    
## household_children          -0.12649    0.06216  -2.035 0.041868 *  
## marital_numeric              0.15975    0.03972   4.022 5.77e-05 ***
## census_msa_numeric          -0.01302    0.04466  -0.292 0.770595    
## age_numeric                  0.46256    0.05945   7.781 7.21e-15 ***
## sex_numeric                  0.20681    0.03482   5.939 2.87e-09 ***
## race_numeric                 0.11283    0.06129   1.841 0.065637 .  
## income_poverty_numeric       0.25982    0.06312   4.116 3.85e-05 ***
## rent_or_own_numeric         -0.01659    0.04484  -0.370 0.711478    
## education_numeric            0.15586    0.05832   2.672 0.007533 ** 
## employment_numeric          -0.11605    0.06079  -1.909 0.056247 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 31391  on 22643  degrees of freedom
## Residual deviance: 22829  on 22617  degrees of freedom
## AIC: 22883
## 
## Number of Fisher Scoring iterations: 4
```

Summary: We see that the following variables are statistically insignificant and therefore, likely, do not significantly contribute to the likelihood of vaccination: behavioral_outside_home, opinion_h1n1_sick_from_vacc, household_adults, census_msa_numeric, race_numeric, rent_or_own_numeric, education_numeric, employment_numeric. We have, however, decided to keep race, education and employment in our subsequent iteration since this is a socio-demographic study.

Validate with the test set

```
logreg01_test <- glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiv
iral_meds + behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands + behavioral_l
arge_gatherings + behavioral_outside_home + doctor_recc_h1n1 + chronic_med_condition + child_u
nder_6_months + health_worker + opinion_h1n1_vacc_effective + opinion_h1n1_risk + opinion_h1n1
_sick_from_vacc + household_adults + household_children + marital_numeric + census_msa_numeric
+ age_numeric + sex_numeric + race_numeric + income_poverty_numeric + rent_or_own_numeric + ed
ucation_numeric + employment_numeric,
    data = prep_test, family = binomial(link = "logit"))

summary(logreg01_test)
```

```
## 
## Call:
## glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + 
##     behavioral_antiviral_meds + behavioral_avoidance + behavioral_face_mask + 
##     behavioral_wash_hands + behavioral_large_gatherings + behavioral_outside_home + 
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months + 
##     health_worker + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + household_adults + household_children +
##     marital_numeric + census_msa_numeric + age_numeric + sex_numeric +
##     race_numeric + income_poverty_numeric + rent_or_own_numeric +
##     education_numeric + employment_numeric, family = binomial(link = "logit"),
##     data = prep_test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4112  -0.5849  -0.3948  -0.1622   3.2491
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.347606   0.316533 -16.894  < 2e-16 ***
## h1n1_concern                -0.177607   0.170844  -1.040 0.298532
## h1n1_knowledge               0.222310   0.149965   1.482 0.138231
## behavioral_antiviral_meds    0.113775   0.185312   0.614 0.539238
## behavioral_avoidance         0.059888   0.106256   0.564 0.573014
## behavioral_face_mask         0.103376   0.153775   0.672 0.501418
## behavioral_wash_hands        0.019636   0.134433   0.146 0.883868
## behavioral_large_gatherings -0.190850   0.093547  -2.040 0.041335 *
## behavioral_outside_home     -0.146840   0.101242  -1.450 0.146952
## doctor_recc_h1n1             1.518047   0.087881  17.274  < 2e-16 ***
## chronic_med_condition        0.154751   0.091381   1.693 0.090367 .
## child_under_6_months         0.315504   0.136864   2.305 0.021153 *
## health_worker                0.705761   0.118111   5.975 2.30e-09 ***
## opinion_h1n1_vacc_effective  2.751675   0.222948  12.342  < 2e-16 ***
## opinion_h1n1_risk            2.041419   0.140470  14.533  < 2e-16 ***
## opinion_h1n1_sick_from_vacc -0.094505   0.132599  -0.713 0.476023
## household_adults             0.028497   0.192992   0.148 0.882611
## household_children           0.042386   0.154803   0.274 0.784234
## marital_numeric              0.143246   0.100544   1.425 0.154241
## census_msa_numeric           0.007934   0.108738   0.073 0.941833
## age_numeric                  0.665874   0.149580   4.452 8.52e-06 ***
## sex_numeric                  0.175212   0.087598   2.000 0.045481 *
## race_numeric                -0.558112   0.164805  -3.386 0.000708 ***
## income_poverty_numeric      -0.116188   0.156699  -0.741 0.458407
## rent_or_own_numeric         -0.069338   0.110518  -0.627 0.530404
## education_numeric            0.096723   0.144737   0.668 0.503963
## employment_numeric           0.154321   0.153807   1.003 0.315696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5236.8  on 4910  degrees of freedom
## Residual deviance: 3890.8  on 4884  degrees of freedom
## AIC: 3944.8
##
## Number of Fisher Scoring iterations: 5
```

*Obtain the predicted values of the target variable for each record in the data set

```
pred = predict(logreg01, newdata=prep_test)
predicted.classes <- factor(ifelse(pred > 0.5, "1", "0"))
accuracy <- table(pred, prep_test[,"h1n1_vaccine"])
sum(diag(accuracy))/sum(accuracy)

## [1] 0.000407249

confusionMatrix(predicted.classes, factor(prep_test$h1n1_vaccine), positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3296  412
##          1  510  693
##
##                Accuracy : 0.8123
##                  95% CI : (0.8011, 0.8231)
##     No Information Rate : 0.775
##     P-Value [Acc > NIR] : 9.704e-11
##
##                   Kappa : 0.4781
##
##  Mcnemar's Test P-Value : 0.001401
##
##             Sensitivity : 0.6271
##             Specificity : 0.8660
##          Pos Pred Value : 0.5761
##          Neg Pred Value : 0.8889
##              Prevalence : 0.2250
##          Detection Rate : 0.1411
##    Detection Prevalence : 0.2450
##       Balanced Accuracy : 0.7466
##
##        'Positive' Class : 1
##
```

```
auc(prep_test$h1n1_vaccine, pred)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.8304
```

Rationalized Logistic Regression

```
#Build and train baseline model with all remaining features
logreg02 <- glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_
meds + behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands + behavioral_large_
gatherings + doctor_recc_h1n1 + chronic_med_condition + child_under_6_months + health_worker +
opinion_h1n1_vacc_effective + opinion_h1n1_risk + household_children + marital_numeric + censu
s_msa_numeric + age_numeric + sex_numeric + race_numeric + income_poverty_numeric + education_
numeric + employment_numeric,
    data = prep_train_rebal, family = binomial(link = "logit"))

summary(logreg02)
```

```
##
## Call:
## glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge +
##     behavioral_antiviral_meds + behavioral_avoidance + behavioral_face_mask +
##     behavioral_wash_hands + behavioral_large_gatherings + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     opinion_h1n1_vacc_effective + opinion_h1n1_risk + household_children +
##     marital_numeric + census_msa_numeric + age_numeric + sex_numeric +
##     race_numeric + income_poverty_numeric + education_numeric +
##     employment_numeric, family = binomial(link = "logit"), data = prep_train_rebal)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.87736  -0.79276  -0.00131   0.78694   2.87578
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -4.14193    0.11183 -37.038  < 2e-16 ***
## h1n1_concern                 -0.27846    0.06457  -4.313 1.61e-05 ***
## h1n1_knowledge                0.34402    0.05869   5.862 4.58e-09 ***
## behavioral_antiviral_meds     0.13516    0.07466   1.810 0.070243 .
## behavioral_avoidance         -0.09393    0.04107  -2.287 0.022181 *
## behavioral_face_mask          0.14126    0.06558   2.154 0.031251 *
## behavioral_wash_hands         0.09426    0.05057   1.864 0.062347 .
## behavioral_large_gatherings -0.24933    0.03754  -6.642 3.09e-11 ***
## doctor_recc_h1n1              1.69929    0.03756  45.242  < 2e-16 ***
## chronic_med_condition         0.14119    0.03727   3.788 0.000152 ***
## child_under_6_months          0.23436    0.05870   3.992 6.54e-05 ***
## health_worker                 1.00700    0.05023  20.049  < 2e-16 ***
## opinion_h1n1_vacc_effective  2.61975    0.08202  31.942  < 2e-16 ***
## opinion_h1n1_risk             1.87860    0.05533  33.951  < 2e-16 ***
## household_children           -0.12357    0.06196  -1.994 0.046126 *
## marital_numeric               0.14307    0.03638   3.933 8.40e-05 ***
## census_msa_numeric           -0.01374    0.04445  -0.309 0.757292
## age_numeric                   0.48349    0.05569   8.681  < 2e-16 ***
## sex_numeric                   0.20842    0.03465   6.015 1.80e-09 ***
## race_numeric                  0.10442    0.06065   1.722 0.085123 .
## income_poverty_numeric        0.26739    0.06152   4.347 1.38e-05 ***
## education_numeric             0.15843    0.05828   2.719 0.006555 **
## employment_numeric           -0.11123    0.06058  -1.836 0.066363 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31391  on 22643  degrees of freedom
## Residual deviance: 22832  on 22621  degrees of freedom
## AIC: 22878
##
## Number of Fisher Scoring iterations: 4
```

Validate with the test set

```
logreg02_test <- glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiv
iral_meds + behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands + behavioral_l
arge_gatherings + doctor_recc_h1n1 + chronic_med_condition + child_under_6_months + health_wor
ker + opinion_h1n1_vacc_effective + opinion_h1n1_risk + household_children + marital_numeric +
census_msa_numeric + age_numeric + sex_numeric + race_numeric + income_poverty_numeric + educa
tion_numeric + employment_numeric,
    data = prep_test, family = binomial(link = "logit"))

summary(logreg02_test)
```

```
##
## Call:
## glm(formula = h1n1_vaccine ~ h1n1_concern + h1n1_knowledge +
##     behavioral_antiviral_meds + behavioral_avoidance + behavioral_face_mask +
##     behavioral_wash_hands + behavioral_large_gatherings + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     opinion_h1n1_vacc_effective + opinion_h1n1_risk + household_children +
##     marital_numeric + census_msa_numeric + age_numeric + sex_numeric +
##     race_numeric + income_poverty_numeric + education_numeric +
##     employment_numeric, family = binomial(link = "logit"), data = prep_test)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4055  -0.5844  -0.3960  -0.1628   3.2919
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.430937   0.297551 -18.252  < 2e-16 ***
## h1n1_concern                -0.218294   0.166539  -1.311 0.189935
## h1n1_knowledge               0.225287   0.149169   1.510 0.130972
## behavioral_antiviral_meds    0.110708   0.185217   0.598 0.550025
## behavioral_avoidance         0.035263   0.104951   0.336 0.736872
## behavioral_face_mask         0.092560   0.153240   0.604 0.545829
## behavioral_wash_hands       -0.024433   0.131001  -0.187 0.852043
## behavioral_large_gatherings -0.214577   0.092336  -2.324 0.020132 *
## doctor_recc_h1n1             1.514504   0.087793  17.251  < 2e-16 ***
## chronic_med_condition        0.154664   0.091263   1.695 0.090132 .
## child_under_6_months         0.312099   0.136779   2.282 0.022502 *
## health_worker                0.696328   0.117942   5.904 3.55e-09 ***
## opinion_h1n1_vacc_effective  2.751648   0.222567  12.363  < 2e-16 ***
## opinion_h1n1_risk            2.003174   0.137083  14.613  < 2e-16 ***
## household_children           0.051293   0.154174   0.333 0.739362
## marital_numeric              0.156990   0.090568   1.733 0.083025 .
## census_msa_numeric           0.003582   0.108130   0.033 0.973571
## age_numeric                  0.682529   0.140907   4.844 1.27e-06 ***
## sex_numeric                  0.189704   0.086891   2.183 0.029019 *
## race_numeric                -0.574209   0.163055  -3.522 0.000429 ***
## income_poverty_numeric      -0.086647   0.152795  -0.567 0.570661
## education_numeric            0.103927   0.144048   0.721 0.470617
## employment_numeric           0.157528   0.153592   1.026 0.305067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5236.8  on 4910  degrees of freedom
## Residual deviance: 3893.9  on 4888  degrees of freedom
## AIC: 3939.9
##
## Number of Fisher Scoring iterations: 5
```

*Obtain the predicted values of the target variable for each record in the data set

```
pred3 = predict(logreg02, newdata=prep_test)
predicted.classes3 <- factor(ifelse(pred3 > 0.5, "1", "0"))
accuracy3 <- table(pred3, prep_test[,"h1n1_vaccine"])
sum(diag(accuracy))/sum(accuracy)

## [1] 0.000407249

auc(prep_test$h1n1_vaccine, pred3)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.8303

confusionMatrix(predicted.classes3, factor(prep_test$h1n1_vaccine), positive = '1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 3295  412
##          1  511  693
## 
##                Accuracy : 0.8121
##                  95% CI : (0.8008, 0.8229)
##     No Information Rate : 0.775
##     P-Value [Acc > NIR] : 1.224e-10
## 
##                   Kappa : 0.4777
## 
##  Mcnemar's Test P-Value : 0.001257
## 
##             Sensitivity : 0.6271
##             Specificity : 0.8657
##          Pos Pred Value : 0.5756
##          Neg Pred Value : 0.8889
##              Prevalence : 0.2250
##          Detection Rate : 0.1411
##    Detection Prevalence : 0.2452
##       Balanced Accuracy : 0.7464
## 
##        'Positive' Class : 1
## 
```

Naives Bayes

```
#Building a naive bayes model

nb1 <- naiveBayes(formula = h1n1_vaccine ~ + 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 + chronic_med_condition +
    child_under_6_months + health_worker + opinion_h1n1_vacc_effective +
    opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + household_adults +
    household_children + marital_numeric + census_msa_numeric +
    age_numeric + sex_numeric + race_numeric + income_poverty_numeric +
    rent_or_own_numeric + education_numeric + employment_numeric, data = prep_train_rebal)

nb1

## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
## 
## A-priori probabilities:
## Y
##   0   1 
## 0.5 0.5 
## 
## Conditional probabilities:
##    h1n1_knowledge
## Y        [,1]      [,2]
##   0 0.6292175 0.2985338
##   1 0.7171878 0.2940591
```

```
##
##    behavioral_antiviral_meds
## Y         [,1]      [,2]
##   0 0.04672319 0.2110547
##   1 0.06668433 0.2494855
##
##    behavioral_avoidance
## Y        [,1]      [,2]
##   0 0.7303480 0.4437986
##   1 0.7701819 0.4207343
##
##    behavioral_face_mask
## Y        [,1]      [,2]
##   0 0.0572337 0.2322989
##   1 0.1034270 0.3045292
##
##    behavioral_wash_hands
## Y        [,1]      [,2]
##   0 0.8201731 0.3840602
##   1 0.8860625 0.3177493
##
##    behavioral_large_gatherings
## Y        [,1]      [,2]
##   0 0.3464052 0.4758452
##   1 0.3672496 0.4820766
##
##    behavioral_outside_home
## Y        [,1]      [,2]
##   0 0.6673733 0.4711749
##   1 0.7403286 0.4384737
##
##    behavioral_touch_face
## Y        [,1]      [,2]
##   0 0.6673733 0.4711749
##   1 0.7403286 0.4384737
##
##    doctor_recc_h1n1
## Y        [,1]      [,2]
##   0 0.1345169 0.3412219
##   1 0.5306483 0.4990818
##
##    chronic_med_condition
## Y        [,1]      [,2]
##   0 0.2565801 0.4367649
##   1 0.3650415 0.4814631
##
##    child_under_6_months
## Y         [,1]      [,2]
##   0 0.07251369 0.2593480
##   1 0.11808868 0.3227273
##
##    health_worker
## Y         [,1]      [,2]
##   0 0.08426073 0.2777907
##   1 0.23467585 0.4238148
##
##    opinion_h1n1_vacc_effective
## Y        [,1]      [,2]
##   0 0.6935612 0.2496240
##   1 0.8528308 0.1877267
##
```

```
##    opinion_h1n1_risk
## Y        [,1]      [,2]
##   0 0.2777999 0.2910021
##   1 0.5424395 0.3355083
## 
##    opinion_h1n1_sick_from_vacc
## Y        [,1]      [,2]
##   0 0.3217850 0.3294960
##   1 0.3844506 0.3605294
## 
##    household_adults
## Y        [,1]      [,2]
##   0 0.3007125 0.2518677
##   1 0.3046870 0.2397253
## 
##    household_children
## Y       [,1]      [,2]
##   0 0.186304 0.3152067
##   1 0.182035 0.3098559
## 
##    marital_numeric
## Y        [,1]      [,2]
##   0 0.5352411 0.4987785
##   1 0.6001590 0.4898871
## 
##    census_msa_numeric
## Y        [,1]      [,2]
##   0 0.5083907 0.3730738
##   1 0.5070217 0.3751974
## 
##    age_numeric
## Y        [,1]      [,2]
##   0 0.5224342 0.3599509
##   1 0.5586469 0.3539352
## 
##    sex_numeric
## Y        [,1]      [,2]
##   0 0.4127363 0.4923479
##   1 0.3827946 0.4860903
## 
##    race_numeric
## Y        [,1]      [,2]
##   0 0.1303068 0.2869537
##   1 0.1233881 0.2885543
## 
##    income_poverty_numeric
## Y        [,1]      [,2]
##   0 0.5879703 0.3099905
##   1 0.6296149 0.3129200
## 
##    rent_or_own_numeric
## Y        [,1]      [,2]
##   0 0.2404169 0.4273555
##   1 0.2101219 0.4074130
## 
##    education_numeric
## Y        [,1]      [,2]
##   0 0.6293941 0.3071838
##   1 0.6545958 0.2831554
## 
##    employment_numeric
## Y        [,1]      [,2]
##   0 0.7510157 0.3050691
##   1 0.7624095 0.2901196
```

Predictions

Confusion Matrix

```
#Create a confusion matrix to evaluate the model.

# Confusion matrix of training set
t.pred1 <-  table(prep_train_rebal$h1n1_vaccine, ypred1)
rownames(t.pred1) <- c("Actual:Not Vaccinated", "Actual: Vaccinated")
colnames(t.pred1) <- c("Predicted: No Vaccine", "Predicted: Vaccinated")
addmargins(A = t.pred1, FUN = list(Total=sum), quiet = TRUE)

##                          ypred1
##                           Predicted: No Vaccine Predicted: Vaccinated Total
##   Actual:Not Vaccinated                    8406                  2916 11322
##   Actual: Vaccinated                       3371                  7951 11322
##   Total                                   11777                 10867 22644

# Confusion matrix for testing data set
t.pred2 <-  table(prep_test$h1n1_vaccine, ypred2)
rownames(t.pred2) <- c("Actual:Not Vaccinated", "Actual: Vaccinated")
colnames(t.pred2) <- c("Predicted: Not Vaccinated", "Predicted: Vaccinated")
addmargins(A =t.pred2, FUN = list(Total=sum), quiet = TRUE)

##                          ypred2
##                           Predicted: Not Vaccinated Predicted: Vaccinated Total
##   Actual:Not Vaccinated                        2795                  1011  3806
##   Actual: Vaccinated                            339                   766  1105
##   Total                                        3134                  1777  4911
```

Evaluate the training model

```
require(caret)

# Convert the data to factor to run evaluations
prep_train_rebal$h1n1_vaccine <- as.factor(prep_train_rebal$h1n1_vaccine)
# verifying the data type
str(prep_train_rebal$h1n1_vaccine)

##  Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...

#Evaluating the confusion matrix of the training set
pred_p <- predict(nb1, newdata=prep_train_rebal)
predicted.class_2 <- factor(ifelse(pred_p > .5, "1","0"))

## Warning in Ops.factor(pred_p, 0.5): '>' not meaningful for factors

accuracy1 <- table(pred_p, prep_train_rebal[,"h1n1_vaccine"])
sum(diag(accuracy1))/sum(accuracy1)

## [1] 0.7223547

#printing the results
confusionMatrix(data=pred_p, factor(prep_train_rebal$h1n1_vaccine), positive = '1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 8406 3371
##          1 2916 7951
##                                           
##                Accuracy : 0.7224          
##                  95% CI : (0.7165, 0.7282)
##     No Information Rate : 0.5             
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.4447          
##                                           
##  Mcnemar's Test P-Value : 1.03e-08        
##                                           
##             Sensitivity : 0.7023          
##             Specificity : 0.7424          
##          Pos Pred Value : 0.7317          
##          Neg Pred Value : 0.7138          
##              Prevalence : 0.5000          
##          Detection Rate : 0.3511          
##    Detection Prevalence : 0.4799          
##       Balanced Accuracy : 0.7224          
##                                           
##        'Positive' Class : 1               
## 
```

```
#Evaluating the confusion matrix of the testing data set
pred_p <- predict(nb1, newdata=prep_test)
accuracy1 <- table(pred_p, prep_test[,"h1n1_vaccine"])
sum(diag(accuracy1))/sum(accuracy1)
```

```
## [1] 0.7251069
```

```
#Printing the results
confusionMatrix(data=pred_p, factor(prep_test$h1n1_vaccine), positive = '1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 2795  339
##          1 1011  766
##                                           
##                Accuracy : 0.7251          
##                  95% CI : (0.7124, 0.7376)
##     No Information Rate : 0.775           
##     P-Value [Acc > NIR] : 1               
##                                           
##                   Kappa : 0.3517          
##                                           
##  Mcnemar's Test P-Value : <2e-16          
##                                           
##             Sensitivity : 0.6932          
##             Specificity : 0.7344          
##          Pos Pred Value : 0.4311          
##          Neg Pred Value : 0.8918          
##              Prevalence : 0.2250          
##          Detection Rate : 0.1560          
##    Detection Prevalence : 0.3618          
##       Balanced Accuracy : 0.7138
##                                            
##        'Positive' Class : 1
## 
```

Improving the model by smoothing with laplace and usekernals

```
str(prep_train_rebal$h1n1_vaccine)

##  Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...

nb2 <- naiveBayes(formula = h1n1_vaccine ~ +h1n1_knowledge + 
    behavioral_antiviral_meds + behavioral_avoidance + behavioral_face_mask +  behavioral_wash
_hands + behavioral_large_gatherings +behavioral_touch_face +doctor_recc_h1n1 + chronic_med_co
ndition + 
    child_under_6_months + health_worker + opinion_h1n1_vacc_effective + 
    opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + household_adults + 
    household_children + marital_numeric + census_msa_numeric + 
    age_numeric + sex_numeric + race_numeric + income_poverty_numeric + 
    rent_or_own_numeric + education_numeric + employment_numeric, data = prep_train_rebal, lap
lace = 1, usekernals = 1)

nb2

## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace, usekernals = 1)
## 
## A-priori probabilities:
## Y
##   0   1 
## 0.5 0.5 
## 
## Conditional probabilities:
##    h1n1_knowledge
## Y        [,1]      [,2]
##   0 0.6292175 0.2985338
##   1 0.7171878 0.2940591
## 
##    behavioral_antiviral_meds
## Y         [,1]      [,2]
##   0 0.04672319 0.2110547
##   1 0.06668433 0.2494855
## 
##    behavioral_avoidance
## Y        [,1]      [,2]
##   0 0.7303480 0.4437986
##   1 0.7701819 0.4207343
## 
##    behavioral_face_mask
## Y        [,1]      [,2]
##   0 0.0572337 0.2322989
##   1 0.1034270 0.3045292
## 
##    behavioral_wash_hands
## Y        [,1]      [,2]
##   0 0.8201731 0.3840602
##   1 0.8860625 0.3177493
```

```
##
##    behavioral_large_gatherings
## Y        [,1]      [,2]
##   0 0.3464052 0.4758452
##   1 0.3672496 0.4820766
##
##    behavioral_touch_face
## Y        [,1]      [,2]
##   0 0.6673733 0.4711749
##   1 0.7403286 0.4384737
##
##    doctor_recc_h1n1
## Y        [,1]      [,2]
##   0 0.1345169 0.3412219
##   1 0.5306483 0.4990818
##
##    chronic_med_condition
## Y        [,1]      [,2]
##   0 0.2565801 0.4367649
##   1 0.3650415 0.4814631
##
##    child_under_6_months
## Y         [,1]      [,2]
##   0 0.07251369 0.2593480
##   1 0.11808868 0.3227273
##
##    health_worker
## Y         [,1]      [,2]
##   0 0.08426073 0.2777907
##   1 0.23467585 0.4238148
##
##    opinion_h1n1_vacc_effective
## Y        [,1]      [,2]
##   0 0.6935612 0.2496240
##   1 0.8528308 0.1877267
##
##    opinion_h1n1_risk
## Y        [,1]      [,2]
##   0 0.2777999 0.2910021
##   1 0.5424395 0.3355083
##
##    opinion_h1n1_sick_from_vacc
## Y        [,1]      [,2]
##   0 0.3217850 0.3294960
##   1 0.3844506 0.3605294
##
##    household_adults
## Y        [,1]      [,2]
##   0 0.3007125 0.2518677
##   1 0.3046870 0.2397253
##
##    household_children
## Y       [,1]      [,2]
##   0 0.186304 0.3152067
##   1 0.182035 0.3098559
##
##    marital_numeric
## Y        [,1]      [,2]
##   0 0.5352411 0.4987785
##   1 0.6001590 0.4898871
##
```

```
##    census_msa_numeric
## Y        [,1]      [,2]
##   0 0.5083907 0.3730738
##   1 0.5070217 0.3751974
##
##    age_numeric
## Y        [,1]      [,2]
##   0 0.5224342 0.3599509
##   1 0.5586469 0.3539352
##
##    sex_numeric
## Y        [,1]      [,2]
##   0 0.4127363 0.4923479
##   1 0.3827946 0.4860903
##
##    race_numeric
## Y        [,1]      [,2]
##   0 0.1303068 0.2869537
##   1 0.1233881 0.2885543
##
##    income_poverty_numeric
## Y        [,1]      [,2]
##   0 0.5879703 0.3099905
##   1 0.6296149 0.3129200
##
##    rent_or_own_numeric
## Y        [,1]      [,2]
##   0 0.2404169 0.4273555
##   1 0.2101219 0.4074130
##
##    education_numeric
## Y        [,1]      [,2]
##   0 0.6293941 0.3071838
##   1 0.6545958 0.2831554
##
##    employment_numeric
## Y        [,1]      [,2]
##   0 0.7510157 0.3050691
##   1 0.7624095 0.2901196
```

Predictions using smoothing and uskernals

```
ypred1 <- predict(nb2, newdata = prep_train_rebal)
(cbind(ypred1, prep_train_rebal))

ypred2 <- predict(nb2, newdata = prep_test)
(cbind(ypred2, prep_test))

# Confusion matrix of training set
t.pred1 <-  table(prep_train_rebal$h1n1_vaccine, ypred1)
rownames(t.pred1) <- c("Actual:Not Vaccinated", "Actual: Vaccinated")
colnames(t.pred1) <- c("Predicted: No Vaccine", "Predicted: Vaccinated")
addmargins(A = t.pred1, FUN = list(Total=sum), quiet = TRUE)

##                         ypred1
##                          Predicted: No Vaccine Predicted: Vaccinated Total
##   Actual:Not Vaccinated                   8437                  2885 11322
##   Actual: Vaccinated                      3325                  7997 11322
##   Total                                  11762                 10882 22644
```

```
# Confusion matrix for testing data set
t.pred2 <-  table(prep_test$h1n1_vaccine, ypred2)
rownames(t.pred2) <- c("Actual:Not Vaccinated", "Actual: Vaccinated")
colnames(t.pred2) <- c("Predicted: Not Vaccinated", "Predicted: Vaccinated")
addmargins(A =t.pred2, FUN = list(Total=sum), quiet = TRUE)

##                          ypred2
##                           Predicted: Not Vaccinated Predicted: Vaccinated Total
##   Actual:Not Vaccinated                        2820                   986  3806
##   Actual: Vaccinated                            333                   772  1105
##   Total                                        3153                  1758  4911
```

Naive Bayes Prediction 2

```
pred_p2 <- predict(nb2, newdata=prep_train_rebal)
predicted.class_p2 <- factor(ifelse(pred_p > .5, "1","0"))

## Warning in Ops.factor(pred_p, 0.5): '>' not meaningful for factors

accuracy2 <- table(pred_p2, prep_train_rebal[,"h1n1_vaccine"])
sum(diag(accuracy1))/sum(accuracy2)

## [1] 0.1572602

#printing the results
confusionMatrix(data=pred_p2, factor(prep_train_rebal$h1n1_vaccine), positive = '1')

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 8437 3325
##          1 2885 7997
##
##                Accuracy : 0.7258
##                  95% CI : (0.7199, 0.7316)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4515
##
##  Mcnemar's Test P-Value : 2.536e-08
##
##             Sensitivity : 0.7063
##             Specificity : 0.7452
##          Pos Pred Value : 0.7349
##          Neg Pred Value : 0.7173
##              Prevalence : 0.5000
##          Detection Rate : 0.3532
##    Detection Prevalence : 0.4806
##       Balanced Accuracy : 0.7258
##
##        'Positive' Class : 1
##

#Evaluating the confusion matrix of the testing data set
pred_p3 <- predict(nb2, newdata=prep_test)

#predicted.class_2 <- factor(ifelse(pred_p > 0.5, "1","0"))
accuracy3 <- table(pred_p3, prep_test[,"h1n1_vaccine"])
sum(diag(accuracy1))/sum(accuracy3)
```

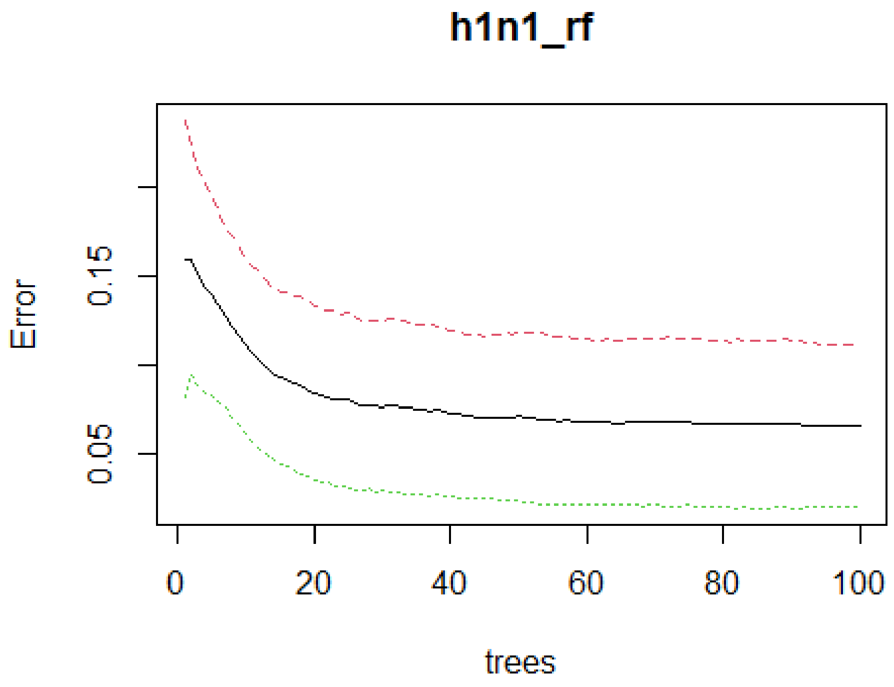
```
## [1] 0.7251069

#Printing the results
confusionMatrix(data=pred_p, factor(prep_test$h1n1_vaccine), positive = '1')

## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 2795  339
##          1 1011  766
## 
##                Accuracy : 0.7251
##                  95% CI : (0.7124, 0.7376)
##     No Information Rate : 0.775
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : 0.3517
## 
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.6932
##             Specificity : 0.7344
##          Pos Pred Value : 0.4311
##          Neg Pred Value : 0.8918
##              Prevalence : 0.2250
##          Detection Rate : 0.1560
##    Detection Prevalence : 0.3618
##       Balanced Accuracy : 0.7138
## 
##        'Positive' Class : 1
## 
```

Random Forest

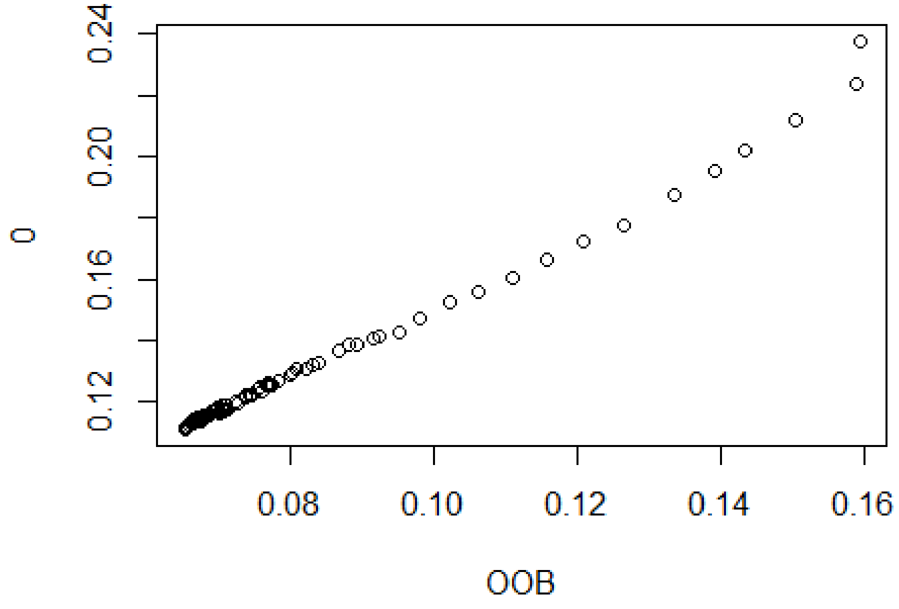
```
prep_train_rebal$h1n1_vaccine <- as.factor(prep_train_rebal$h1n1_vaccine)
prep_test$h1n1_vaccine <-as.factor(prep_test$h1n1_vaccine)
# had to convert the target variables to factors before running the code to run it as a classi
fication model instead of regression.
library(randomForest)
h1n1_rf <- randomForest(h1n1_vaccine ~., data = prep_train_rebal, ntree = 100,proximity = TRUE
)
print(h1n1_rf)

## 
## Call:
##  randomForest(formula = h1n1_vaccine ~ ., data = prep_train_rebal,      ntree = 100, proxim
ity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 5
## 
##         OOB estimate of  error rate: 6.58%
## Confusion matrix:
##       0     1 class.error
## 0 10059  1263  0.11155273
## 1   226 11096  0.01996114
```

the out-of-bag error rate is 6.58%. Meaning that 93.42% of the data was predicted correctly.

```
plot(h1n1_rf)
```



```
plot(h1n1_rf$err.rate)
```



The Mean Decrease Gini below

```
h1n1_rf$importance
```

```
##                              MeanDecreaseGini
## h1n1_concern                        454.9294
## h1n1_knowledge                      327.7793
## behavioral_antiviral_meds            99.1327
## behavioral_avoidance                218.2128
## behavioral_face_mask                106.3110
## behavioral_wash_hands               145.3943
## behavioral_large_gatherings         244.6227
## behavioral_outside_home             142.8629
## behavioral_touch_face               146.0614
## doctor_recc_h1n1                   1463.4372
## chronic_med_condition               219.2353
## child_under_6_months                124.8000
## health_worker                       318.0590
## opinion_h1n1_vacc_effective        1048.2797
## opinion_h1n1_risk                  1174.6974
## opinion_h1n1_sick_from_vacc         488.0349
## household_adults                    393.9290
## household_children                  362.9019
## marital_numeric                     218.5950
## census_msa_numeric                  452.5672
## age_numeric                         593.6944
## sex_numeric                         272.2799
## race_numeric                        288.9682
## income_poverty_numeric              333.7064
## rent_or_own_numeric                 191.1977
## education_numeric                   481.6087
## employment_numeric                  309.3920
```

what is mtry?

```
h1n1_rf$mtry
```

```
## [1] 5
```

why the accuracy?

```
p1<- predict(h1n1_rf, prep_train_rebal)
confusionMatrix(p1, prep_train_rebal$h1n1_vaccine, positive = '1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction     0     1
##          0 11264    16
##          1    58 11306
## 
##                Accuracy : 0.9967
##                  95% CI : (0.9959, 0.9974)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.9935
## 
##  Mcnemar's Test P-Value : 1.878e-06
## 
##             Sensitivity : 0.9986
##             Specificity : 0.9949
##          Pos Pred Value : 0.9949
##          Neg Pred Value : 0.9986
##              Prevalence : 0.5000
##          Detection Rate : 0.4993
##    Detection Prevalence : 0.5019
##       Balanced Accuracy : 0.9967
##
##        'Positive' Class : 1
##
```

prepare new model based on the gini coefficiency? what does it say?

```
h1n1_meangini <- randomForest(h1n1_vaccine ~ doctor_recc_h1n1+
opinion_h1n1_risk + opinion_h1n1_vacc_effective +
age_numeric + opinion_h1n1_sick_from_vacc + education_numeric +
h1n1_concern + census_msa_numeric + household_adults + household_children + income_poverty_num
eric + h1n1_knowledge +
health_worker + employment_numeric +race_numeric+
sex_numeric, data = prep_train_rebal, ntree = 100,proximity = TRUE)
```

gini model confusion matrix shows that the model accuracy decreased using the Mean Decrease Gini

```
pred_gini<- predict(h1n1_meangini, prep_train_rebal)
confusionMatrix(pred_gini, prep_train_rebal$h1n1_vaccine, positive = '1')

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 11067   102
##          1   255 11220
##
##                Accuracy : 0.9842
##                  95% CI : (0.9825, 0.9858)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9685
##
##  Mcnemar's Test P-Value : 8.646e-16
##
##             Sensitivity : 0.9910
##             Specificity : 0.9775
##          Pos Pred Value : 0.9778
##          Neg Pred Value : 0.9909
##              Prevalence : 0.5000
##          Detection Rate : 0.4955
##    Detection Prevalence : 0.5068
##       Balanced Accuracy : 0.9842
##
##        'Positive' Class : 1
##
```

Random Forest Model validation

```
p2<- predict(h1n1_rf, prep_test)
confusionMatrix(p2, prep_test$h1n1_vaccine, positive = '1')

## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction    0    1
##          0 3376  489
##          1  430  616
##                                           
##                Accuracy : 0.8129          
##                  95% CI : (0.8017, 0.8237)
##     No Information Rate : 0.775           
##     P-Value [Acc > NIR] : 4.802e-11       
##                                           
##                   Kappa : 0.4531          
##                                           
##  Mcnemar's Test P-Value : 0.05572         
##                                           
##             Sensitivity : 0.5575          
##             Specificity : 0.8870          
##          Pos Pred Value : 0.5889          
##          Neg Pred Value : 0.8735          
##              Prevalence : 0.2250          
##          Detection Rate : 0.1254          
##    Detection Prevalence : 0.2130          
##       Balanced Accuracy : 0.7222          
##                                           
##        'Positive' Class : 1               
##
```

Gini Model Decrease Validation

```
pred_gini_test<- predict(h1n1_meangini, prep_test)
confusionMatrix(pred_gini_test, prep_test$h1n1_vaccine, positive = '1')

## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 3287  455
##          1  519  650
##                                           
##                Accuracy : 0.8017          
##                  95% CI : (0.7902, 0.8127)
##     No Information Rate : 0.775           
##     P-Value [Acc > NIR] : 3.057e-06       
##                                           
##                   Kappa : 0.4428          
##                                           
##  Mcnemar's Test P-Value : 0.04352         
##                                           
##             Sensitivity : 0.5882          
##             Specificity : 0.8636          
##          Pos Pred Value : 0.5560          
##          Neg Pred Value : 0.8784          
##              Prevalence : 0.2250          
##          Detection Rate : 0.1324          
##    Detection Prevalence : 0.2380          
##       Balanced Accuracy : 0.7259          
##                                           
##        'Positive' Class : 1               
##
```