GROUP 7 DATA REPORT ON H1N1 AND SEASONAL VACCINE

GROUP MEMBERS; **JULIA KARANJA**

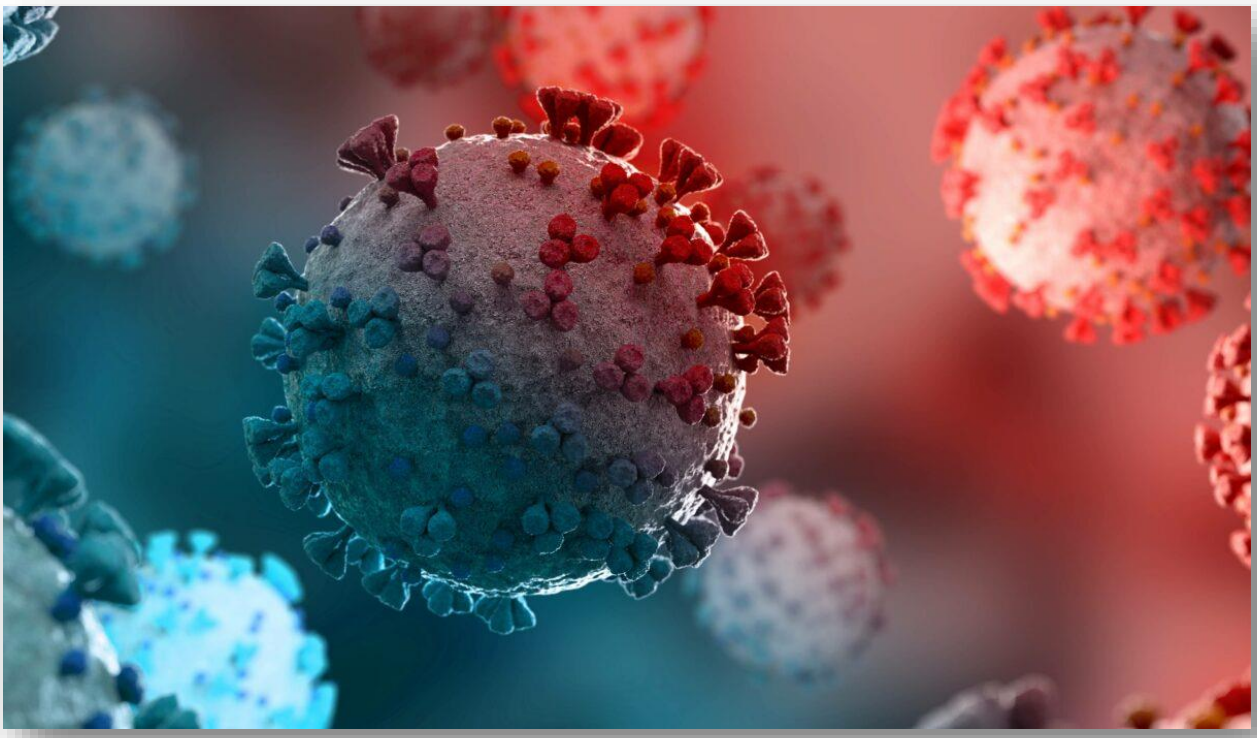
PRIDE AMOS

DANIEL KIMUTAI

CALVINCE OCHIENG

BELINDA NYAMAI

DATASET; https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/



1. BUSINESS UNDERSTANDING

1.1 OVERVIEW

Influenza, commonly known as "the flu", is an infectious disease caused by *influenza viruses*. Symptoms range from mild to severe and often include fever, runny nose, sore throat, muscle pain, headache, coughing, and fatigue. These symptoms begin from one to four days after exposure to the virus (typically two days) and last for about 2–8 days. Diarrhea and vomiting can occur, particularly in children. Influenza may progress to pneumonia, which can be caused by the virus or by a subsequent bacterial infection. Other complications of infection include acute respiratory distress syndrome, meningitis, encephalitis, and worsening of pre-existing health problems such as asthma and cardiovascular disease.

There are four types of influenza virus, termed influenza viruses A, B, C, and D. *Influenza A virus* (IAV) and *Influenza B virus* (IBV) primarily affect humans. They circulate in humans and cause seasonal epidemics of disease known as flu season almost every winter in the United States. IAV are the most common influenza viruses known to cause global epidemics of flu disease.

According to the World Health Organization, people such as those aged 65 years and older, young children and people with certain health conditions are at a higher risk of serious flu complications. For the influenza A and B viruses that routinely spread in people, human influenza viruses are responsible. Most experts believe that in humans, influenza viruses are primarily transmitted through respiratory droplets produced from coughing and sneezing. Less often, a person might get flu by touching a surface or object that has flu droplets on it and touching their own mouths, nose or possibly their eyes. The best way to reduce the risk of flu and its serious complications is by getting vaccinated each year.

1.2 PROBLEM STATEMENT

The influenza virus is constantly mutating by essentially putting on ever changing disguises to penetrate immune systems. A pandemic is caused by a new virus that emerges and easily infects people to when they have no immunity. The recent COVID-19 vaccine had dire consequences to the health sector and the economy at large where many people lost their lives and major businesses experienced losses. Influenza vaccinations are therefore important to protect people who are vulnerable to contracting the disease like young children, older people, pregnant women and people with vulnerable immune systems. This project aims towards understanding how people's background, opinions and health behaviours are related to their decisions to get the H1N1 and seasonal flu vaccines.

1.3 PROPOSED SOLUTION

The expectation is that this project will provide insight to health stakeholders on the major factors they should consider when conducting vaccination campaigns and increase vaccination coverage among high risk groups. The project will also strengthen national, regional and global influenza response capacities including; diagnostics, susceptibility monitoring, disease surveillance and outbreak responses.

1.4 SPECIFIC OBJECTIVES

1. To determine how people's backgrounds like; age, education, race, sex, marital status, employment affects their decisions to get H1N1 or Seasonal flu vaccines.

2. To determine how people's opinions on H1N1 vaccine and seasonal flu vaccine affect their decision to get vaccinated.

3. To determine how Health behaviours like; washing hands, buying face masks, avoiding close contact with others, taking antiviral medication affect the decision to get H1N1 or Seasonal flu vaccines.

1.5 RESEARCH QUESTION

Does the background of people, opinions on influenza vaccines and health behaviours have any impact on obtaining the vaccine?

2. DATA UNDERSTANDING

2.1 EXPLORING DATA

The data set includes fields that represent features influencing people to undertake the H1N1 and Seasonal vaccine. The dataset contains 26707 rows and 38 columns with no duplicates. The column definitions are displayed below. Some columns were found irrelevant for our analysis hence they were dropped.

Column Name	Description
h1n1_vaccine	Whether respondent received H1N1 flu vaccine
seasonal_vaccine	Whether respondent received seasonal flu vaccine.
respondent_id	Unique and random identifier.
h1n1_concern	Level of concern about the H1N1 flu. (0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.)
h1n1_knowledge	Level of knowledge about H1N1 flu. (0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.)
behavioral_antiviral_meds	Has taken antiviral medications. (binary)
behavioral_avoidance	Has avoided close contact with others with flu-like symptoms. (binary)
behavioral_face_mask	Has bought a face mask. (binary)
behavioral_wash_hands	Has frequently washed hands or used hand sanitizer. (binary)
behavioral_large_gatherings	Has reduced time at large gatherings. (binary)
behavioral_outside_home	Has reduced contact with people outside of own household. (binary)
behavioral_touch_face	Has avoided touching eyes, nose, or mouth. (binary)
doctor_recc_h1n1	H1N1 flu vaccine was recommended by doctor. (binary)
doctor_recc_seasonal	Seasonal flu vaccine was recommended by doctor. (binary)
chronic_med_condition	Has any of the following chronic medical conditions: asthma or another lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines

	taken for a chronic illness. (binary)
child_under_6_months	Has regular close contact with a child under the age of six months. (binary)
health_worker	Is a healthcare worker. (binary)
health_insurance	Has health insurance. (binary)
opinion_h1n1_vacc_effective	Respondent's opinion about H1N1 vaccine effectiveness. (1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.)
opinion_h1n1_risk	Respondent's opinion about risk of getting sick with flu without vaccine. (1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.)
opinion_h1n1_sick_from_vacc	Respondent's worry of getting sick from taking H1N1 vaccine. (1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.)
opinion_seas_vacc_effective	Respondent's opinion about seasonal flu vaccine effectiveness. (1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.)
opinion_seas_risk	Respondent's opinion about risk of getting sick with seasonal flu without vaccine. (1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.)
opinion_seas_sick_from_vacc	Respondent's worry of getting sick from taking seasonal flu vaccine. (1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.)
age_group	Age group of respondent.
education	Self-reported education level.
race	Race of respondent.
sex	Sex of respondent.
income_poverty	Household annual income of respondent with respect to 2008 Census poverty thresholds.
marital_status	Marital status of respondent.
rent_or_own	Housing situation of respondent.
employment_status	Employment status of respondent.
hhs_geo_region	Respondent's residence using a 10 region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
census_msa	Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
household_adults	Number of other adults in household, top-coded to 3.
household_children	Number of children in household, top-coded to 3.
employment_industry	Type of industry respondent is employed in. Values are represented as short random character strings.
employment_occupation	Type of occupation of respondent. Values are represented as short random character strings.

2.2 DATA PREPARATION

The following steps were followed in preparing the data;

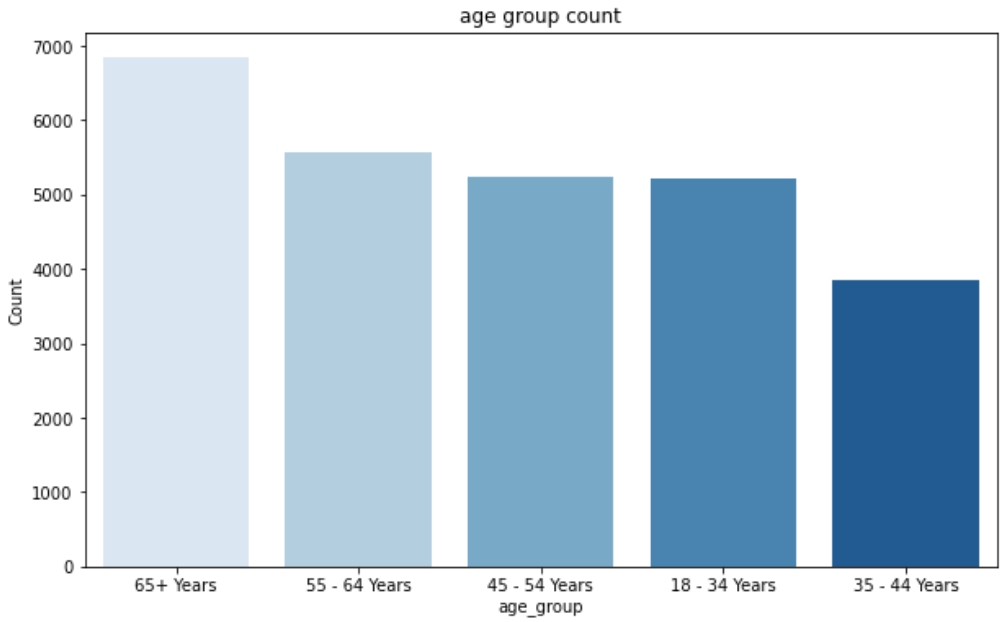
- Importing the necessary libraries
- Loading the dataset from the CSV format it was stored in
- Creating a new data frame with the necessary columns for our research
- Cleaning the data
  - Checking for missing values and replacing them with mode
  - Renaming the geographical region column values

3. EXPLORATORY DATA ANALYSIS

The data at hand represents information about respondents concerning their backgrounds, opinions, and health behaviors.

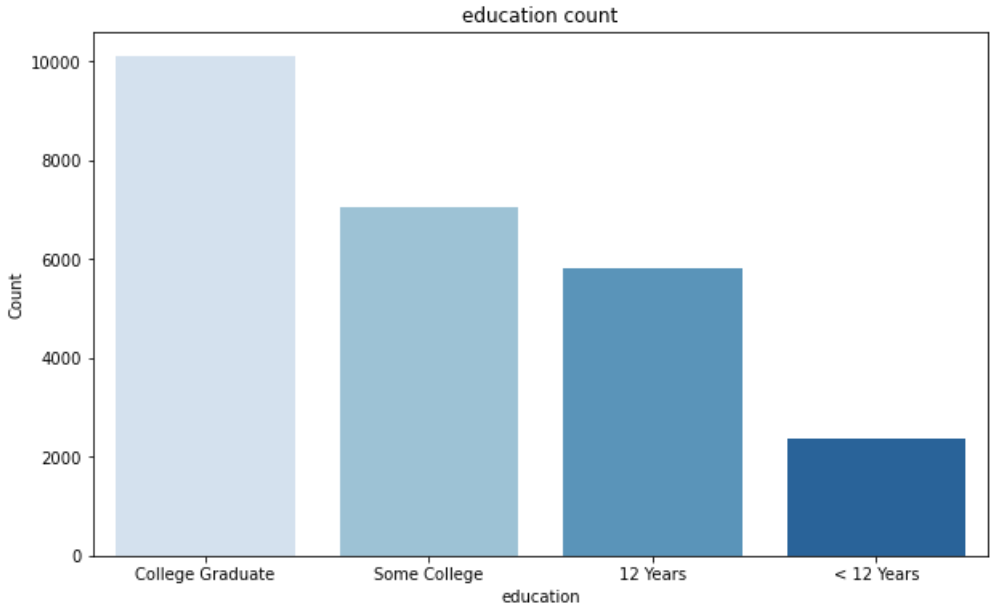
3.1 UNIVARIATE DATA ANALYSIS

Representation of age groups in the dataset



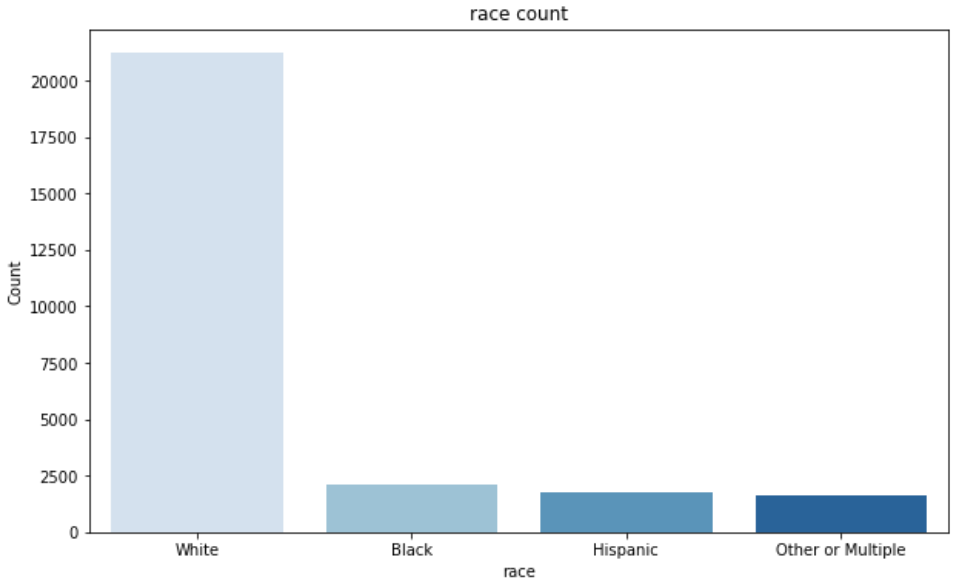
The highest age group is 65 years and above while the lowest age group ranges between 35 and 44 years.

Representation of the various education levels



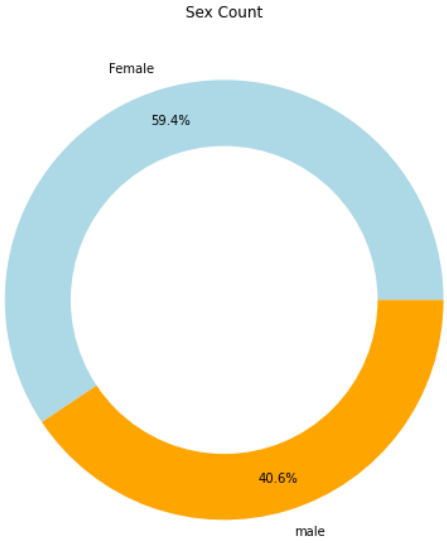
There are a high number of college graduates compared to people with less than 12years education level.

Representation of race



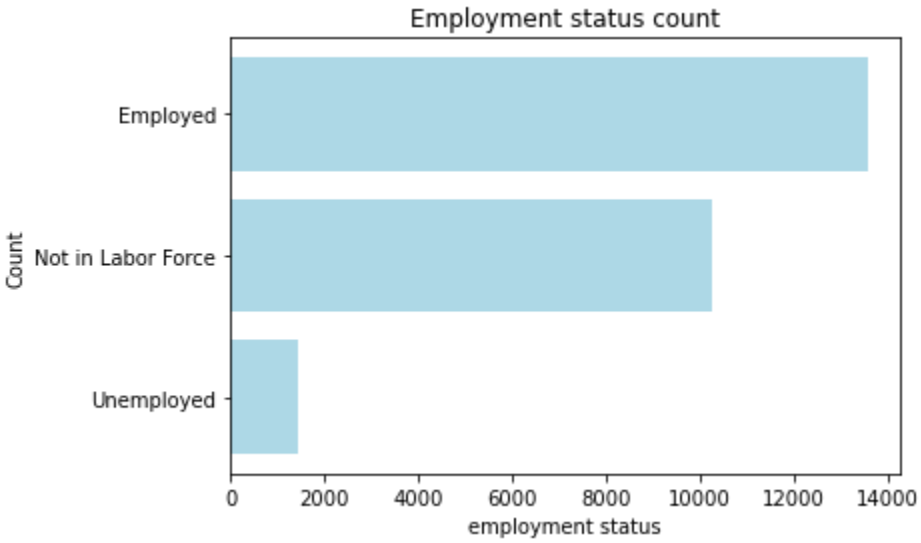
There is a high number of white people and a low number of people belonging to other or multiple races.

Representation of gender



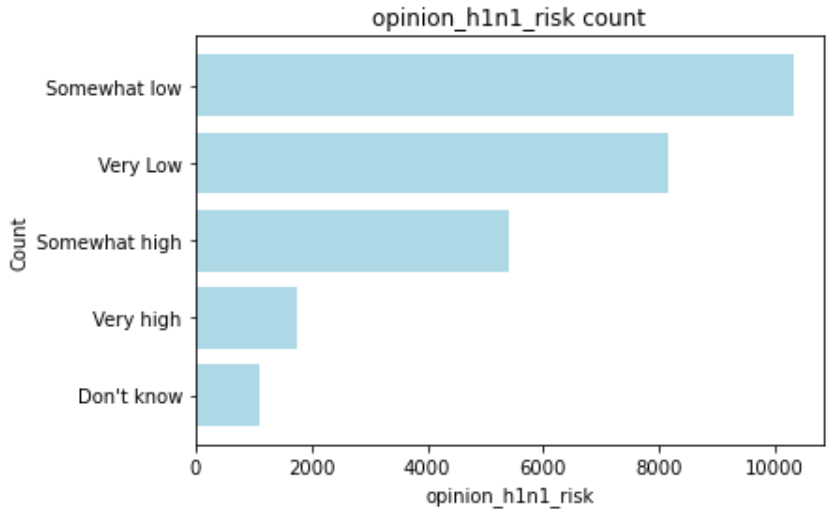
The dataset contains a high number of females compared to males.

Representation of employment status



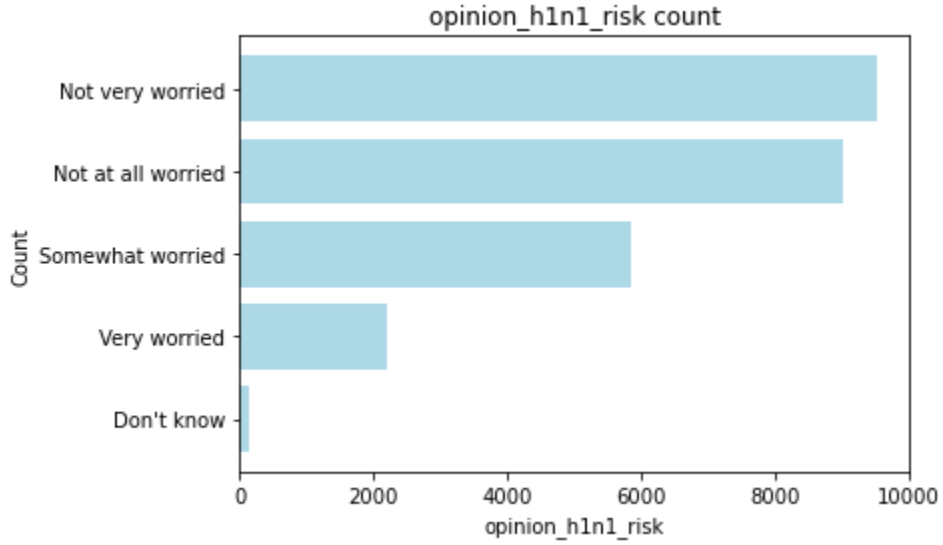
The employed people are the majority in the dataset compared to the unemployed who are few.

Representation of people's opinion of getting the flu without vaccination



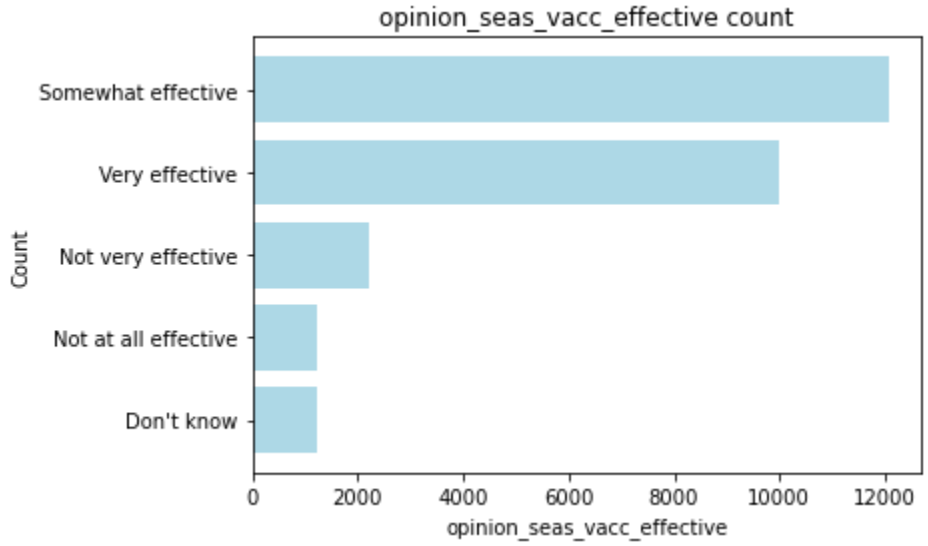
Most people believe there are low chances of getting the flu even if they fail to obtain the vaccine.

Representation of how worried people are after getting the H1N1 vaccine



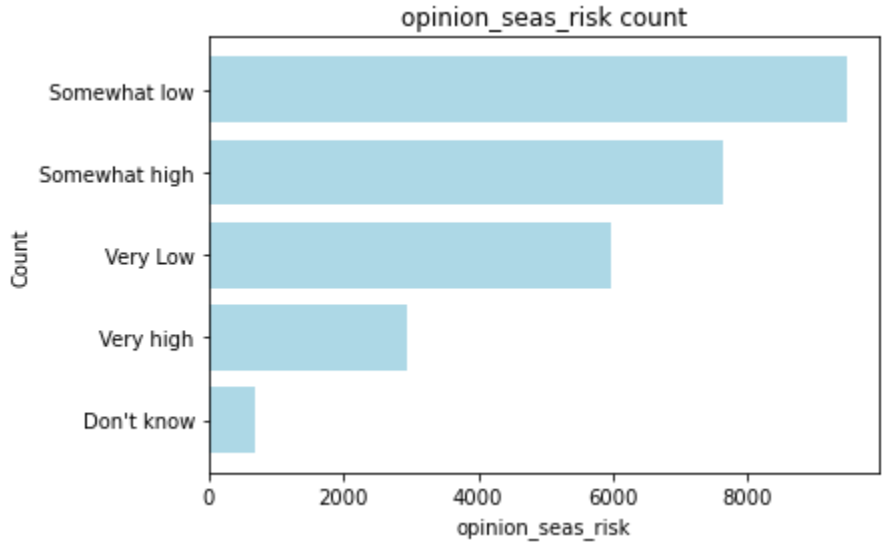
Most people are not worried about contracting the flu or the side effects associated with the vaccine after obtaining the H1N1 vaccine.

Representation of people's opinion about seasonal flu vaccine effectiveness.



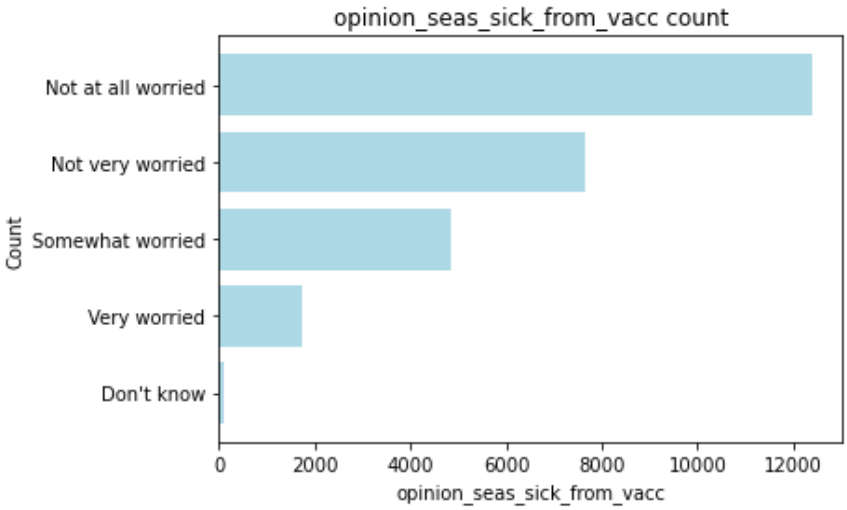
The most popular opinion from people is that the seasonal vaccine is somewhat effective.

Representation of people's worry of getting sick without taking the seasonal flu vaccine.



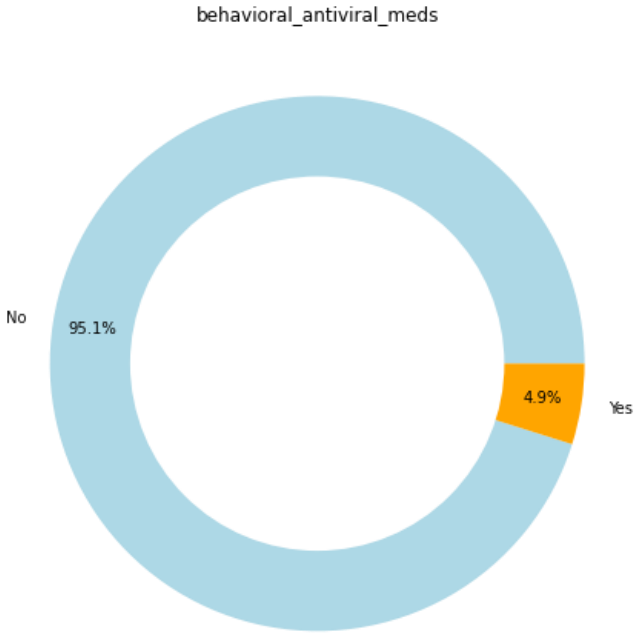
Most people are not worried about contracting seasonal flu even though they do not get the seasonal flu vaccine.

Representation of people's worry of getting sick after taking the Seasonal vaccine.



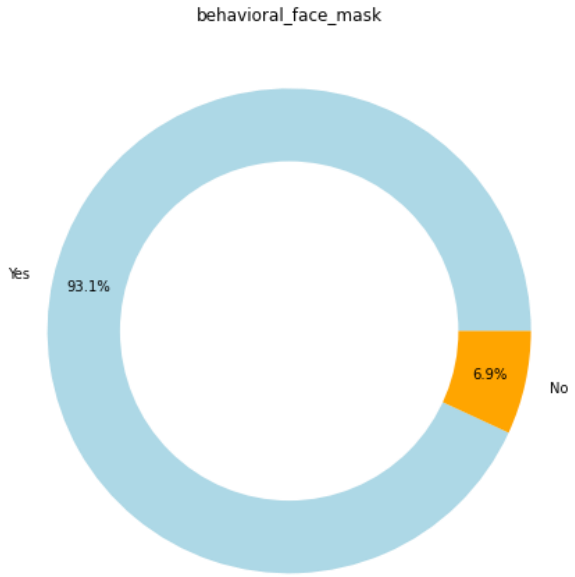
Most people are not worried at all about contracting the flu after taking the Seasonal vaccine.

Representation of people's behaviour to antiviral meds



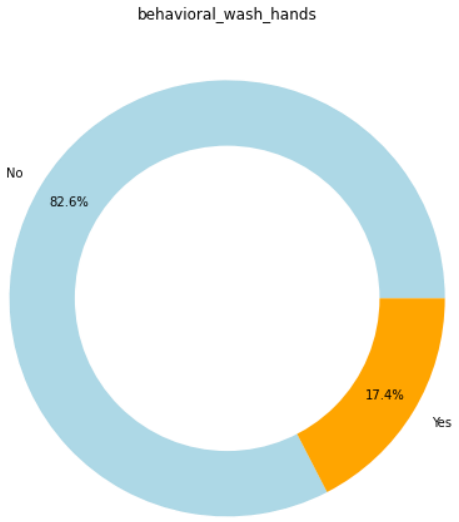
Majority of people rarely take antiviral medicines as a prevention measure for contacting the flu.

Representation of people's behaviour towards wearing a face mask.



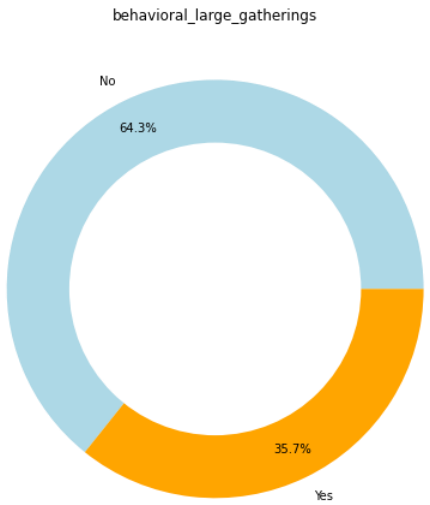
Majority of the population embrace the idea of wearing face masks as prevention measure against flu.

Representation of people's behaviour of washing hands



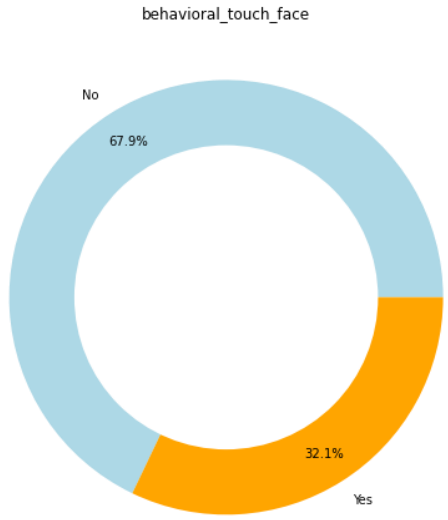
Majority of people have not embraced the idea of washing their hands as a prevention measure against flu.

Representation of people's behaviour of avoiding large gatherings



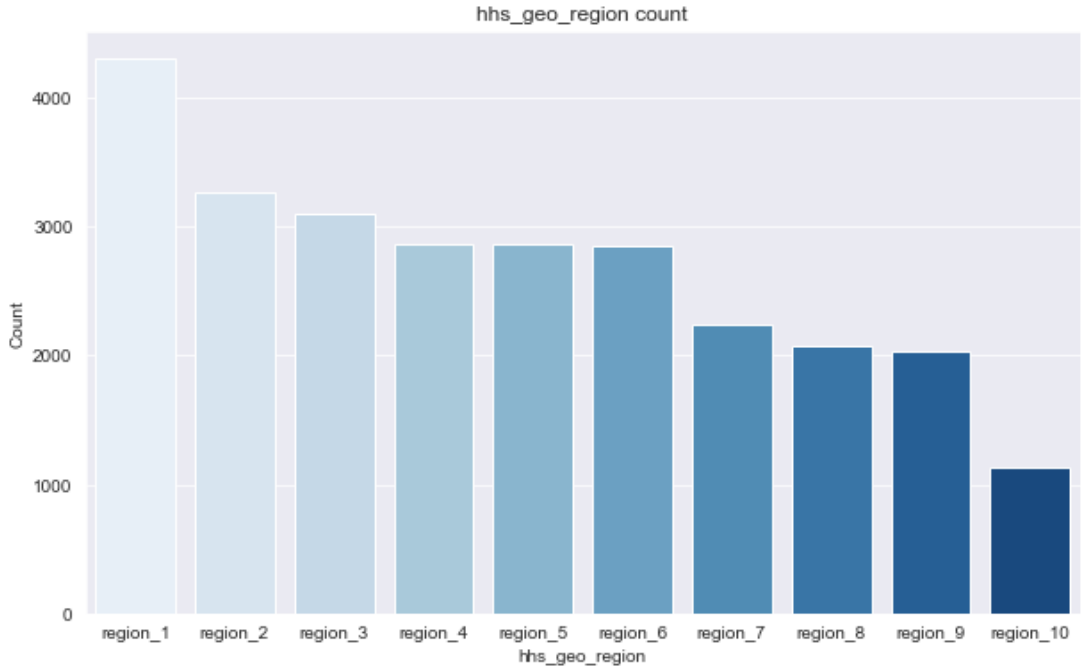
Majority of people have not embraced the idea of avoiding large gatherings as a prevention measure against flu.

Representation of people's behaviour of touching their faces



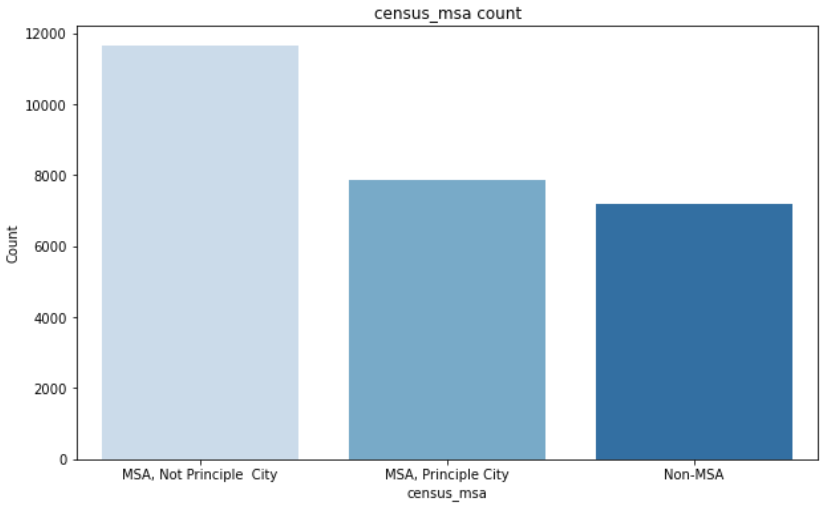
Most people have not embraced the idea of failing to touch their face as a prevention measure against flu.

Representation of respondent's residence using a 10 region geographic classification



Region 1 is the most populated region while region 10 has the least population

Representation of respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.

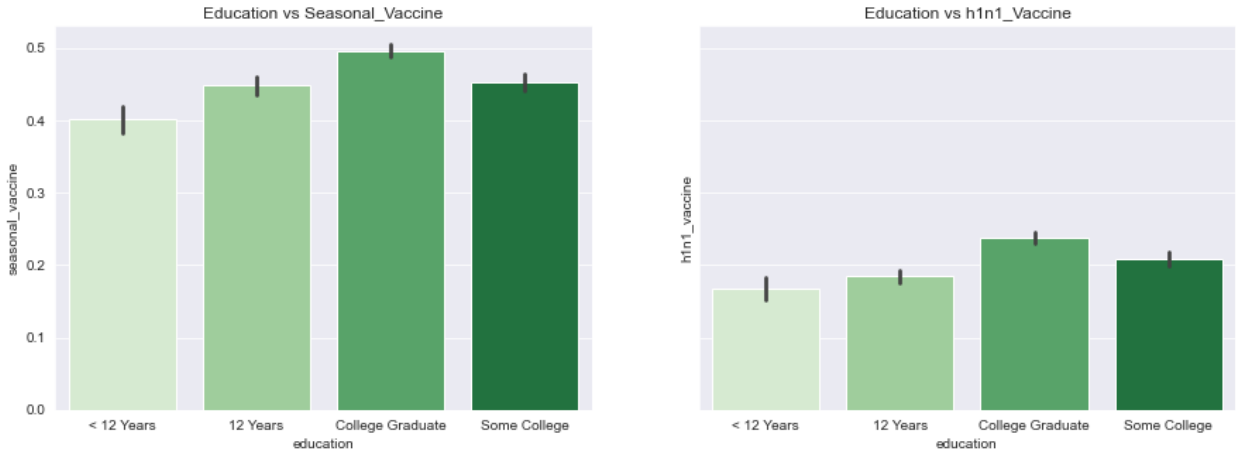


The metropolitan area that is not a principle city had the highest population.

3.2 BIVARIATE DATA ANALYSIS

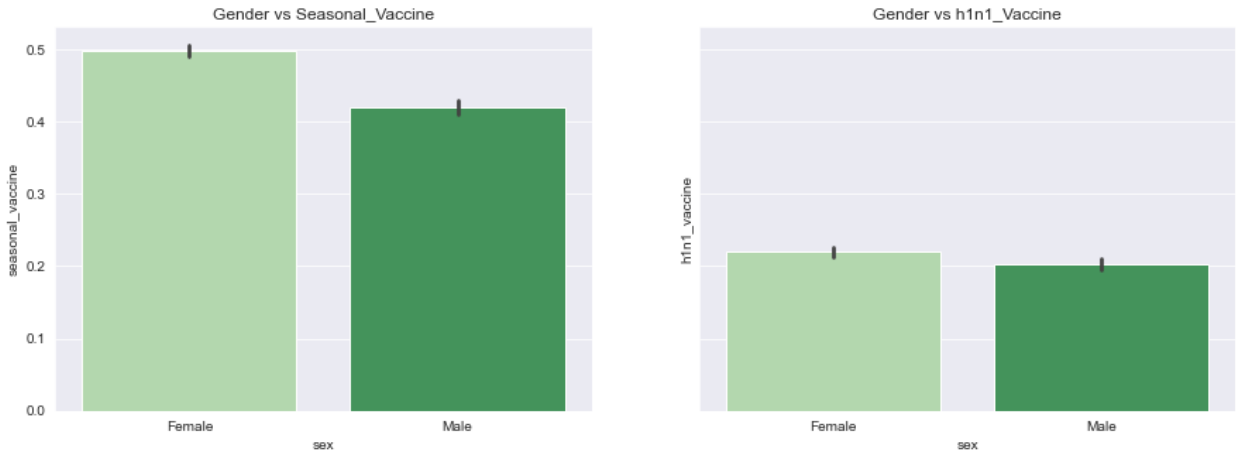
To check for the relationship of various factors with H1N1 and Seasonal vaccine; education, gender, employment status and age group were plotted against the two vaccines.

Displaying relationship between H1N1 and Seasonal vaccine and various educational levels



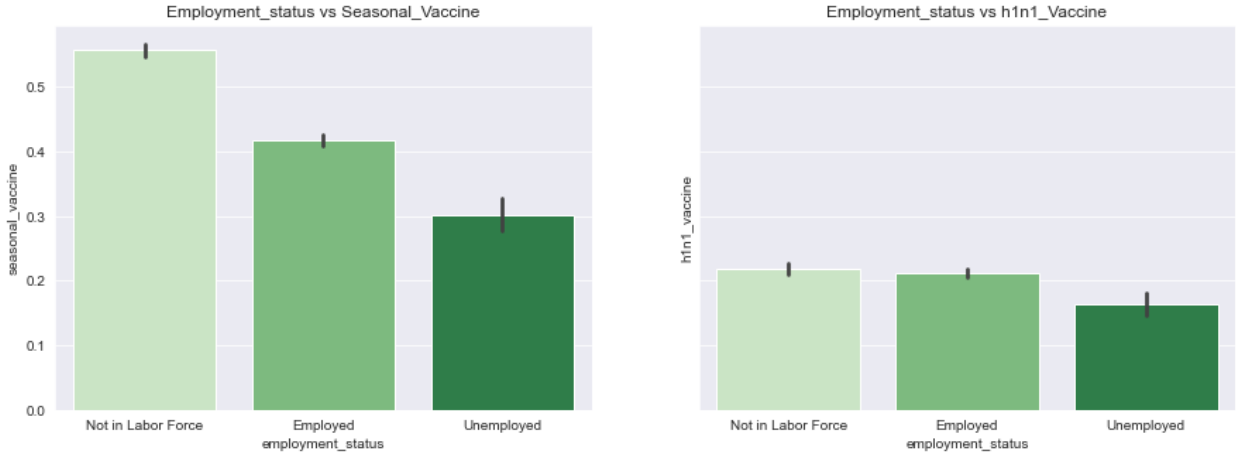
Most of the people who took both of the vaccines were college graduates, this shows that education level has a huge impact on obtaining a certain vaccine. Evidently, those with low education level of less than 12years had a low turnout in obtaining any of the vaccines.

Displaying relationship between H1N1 and Seasonal vaccine and gender



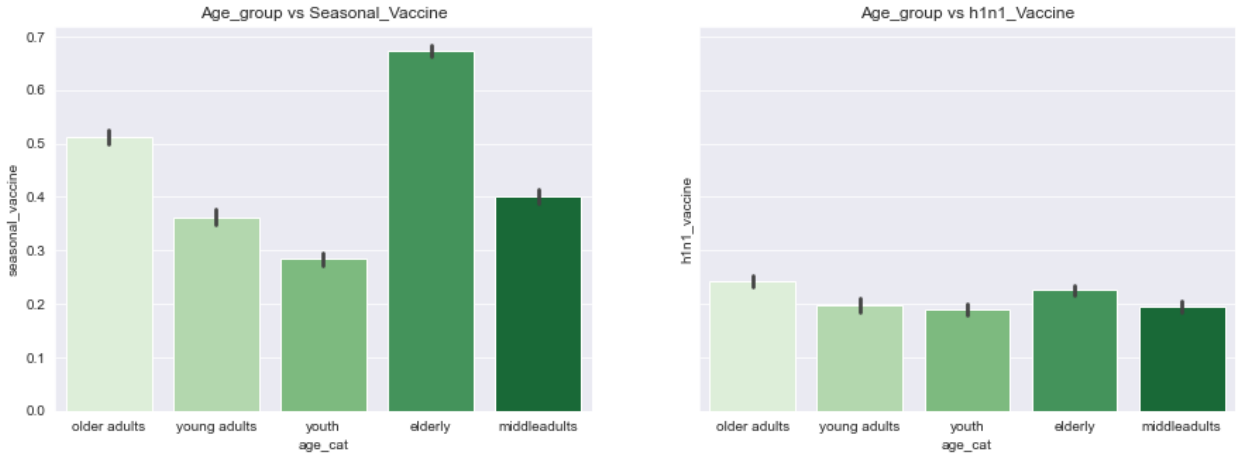
With regard to gender, females had a high turnout in obtaining both of the vaccines compared to males. This could be attributed to the fact that pregnant women are also at a high risk of contacting the flu. When pregnant women obtain the vaccine, this adds on to the number of women who take the vaccine compared to men.

Displaying relationship between H1N1 and Seasonal vaccine and employment status



Those not in the labour force have had a huge turnout of taking the vaccines compared to those who are employed. This evidently shows that the low turnout for those employed may be due to assurance of finances in the case illness occurs contrary to those not in the labour force.

Displaying relationship between H1N1 and Seasonal vaccine and age group



The elderly mainly obtain vaccination compared to the rest of the age groups. This could be attributed to the low immune of the elderly people hence the high risk of flu transmission. The rest of the age groups comprising of young adults have a relatively stronger immune making them less prone to contracting the flu illness.

4. DATA PREPROCESSING

This part of the analysis involves looking at trying to find solutions for:

- missing values and all those values that are falsely labeled, but are missing values and some of the rather strangely looking labels
- Outliers and noise
- Encoding of categorical data

Training the model

The model is trained with the training features (X_train) and training labels (y_train) and is given some new data it hasn't seen before (X_test) to evaluate how well it classifies the new data. The training or test split percentages do not however affect our workflow. 70% of the data was set for training purposes while 30% was set for testing purposes.

MinMaxScaler

This is a technique used to transform features by scaling each feature to a given range. It is an estimator that scales and translates each feature individually such that it is in the given range on the training set.

One Hot Encoding

Most of the data has already been encoded; hence the answers are represented as numbers 0 or 1 (for binary variables) or form 0 to 5 (for multiclass variables). However there are a few categories left that have to be encoded. Since the variables don't have a high cardinality, **one-hot encoding** (OHE) is used to create a binary variable for each label of a categorical variable. We will end up with n-1 new feature for each original feature, where n represents the number of labels of the feature.

Feature Importance

It is a technique that calculates a score for all the input features for a given model. A higher score means that the specific feature will have a larger effect on the model used to predict a certain variable. It is essentially used for data understanding, model improvement and model interpretability.

Pipeline with StandardScaler and KNeighbours classifier

The basic pipeline implemented consists of StandardScaler for pre-processing data and KNeighborsClassifier that strives to find patterns to accurately map inputs to outputs based on known ground truths.

5. EVALUATION

We used different models to come up with a successful predictions, our success metrics was based on the accuracy score of above 65% or an A_U_C score of above 70%.Listed below are the various models that we used and their accuracy score

- KNeighborsClassifier with an accuracy score of 56.59%
- Random forest classifier with an accuracy score of 64.00%
- XG boost with an accuracy score of 66%
- BinaryRelevance(LogisticRegression) classifier with an accuracy score of 67%
- BinaryRelevance guassian naive bayes with an accuracy score of 58.75%
- Multioutput classifier with an average AUC of 83.73%

Hence we decided that the best model was the Multioutput classifier with an average AUC of 83.73 %.

6. RECOMMENDATIONS

Since vaccination is the main preventive strategy for influenza, optimizing formations and identifying factors that interfere with the administration of the vaccine is vital. Identifying factors that produce a priming effect and enhance response is important in understanding how to improve efficiency of influenza vaccine. Prospective safety monitoring followed by rigorous signal refinement is critical to inform decision making by regulatory and public health agencies.

7. CONCLUSION

Our study demonstrates that college graduates, females, those not in the labor force and the elderly have the highest turn out in obtaining both the h1n1 vaccine and the seasonal vaccine. Evidently, the probability of an individual to get vaccinated against flu is often dictated by; demographic factors, what people perceive and day to day behaviors towards preventing infection. Generally, the model indicates that; younger age-groups are less likely to get vaccinated, individuals belonging to black race are less likely to get vaccinated and people who rate higher at falling sick are likely to get vaccinated.

8. LIMITATIONS

Vaccination data was obtained from electronic records hence subject to errors and misclassification.