Influenza Influences: Prediction of Influenza Vaccine Distribution Using Search Engine Query Data

G2 - The InFLUenzers

Data Science Capstone Project
Data Acquisition and Pre-Processing Report

Date:

May 8, 2022

Team Members:

Name: Andrew Chen

Name: Jackie Glosson

Name: Tien Nguyen

Name: Ashley Wheeler

Identifying Data

Data Sources:

- Google Trends
    a. Source: direct download from google trends

We selected Google trends as a data source because google trends is a free tool derived from the google search engine that would allow us to understand what people are interested in knowing about the flu shot (in our case) in real-time. We can use the data from years back to understand and gain insight into the general population's behavior.

- CDC Influenza Vaccination Coverage for All Ages (6+ Months)
    a. Source: Direct download from the CDC

Having decided that we wanted to investigate the flu shot seasonality, we thought it would be good to collect data on influenza vaccination coverage for all ages in the US, categorized by age group and race/ethnicity. The CDC had data available on influenza vaccination coverage for the US. The CDC website provides direct access to important health and safety topics, scientific articles, data and statistics, and tools and resources.

- Demographic Data
    a. Age Distribution over Sixty-Five
        i. Source: census Age
    b. Median Household Income
        i. Source: census Income
    c. Education Attainment
        i. Source: census Education
    d. Racial Composition
        i. Source: census Race
    e. Health Insurance
        Source: Insurance

These data sources were chosen because when we were in the process of our literature search, we found that there are direct factors that influence flu vaccination. These factors include adults in older age groups (75+ years old), adults with a bachelor's degree or higher education level, racial composition, median household income, and whether they have access to health insurance (Abbas et al., 2018).

- Weather:
    a. Source: weather
    b. The data records the average temperature for each month for each state

We chose to use weather data to show the seasonality of the flu season and people taking the flu shot. Due to temperature, the flu tends to spike in the fall and winter (Larson, 2021). Therefore, weather data would be a valuable factor to observe.

- Political Affiliation:
    a. Source: Political Affiliation

Recently, covid vaccination has become a partisan issue, and stark differences in vaccination rates are observed between those identifying with different political parties. Recently this effect has been observed in flu shot distribution (Enten, 2021). Therefore, we will examine political affiliation as a factor. A CNN article stated, "It seems plausible that the push to get the Covid-19 vaccine has led to more Democrats getting the flu shot, while it has had the opposite effect on Republicans" (Enten, 2021). Since COVID-19, we may see more data on the relationship between political affiliation and vaccines, so we will explore political affiliation as a possible factor.

Acquisition Process:

- Google Trends
    a. Source: direct download from google trends
    b. Ten (10) datasets, each corresponding to ten different states: Massachusetts, Connecticut, Maryland, Pennsylvania, Minnesota, Mississippi, Idaho, Wyoming, Louisiana, and Montana (These states were chosen due to being the top 5 and bottom 5 for the term searches that we are using, i.e., search term: "flu shot")
    c. Each dataset contains monthly data, beginning from April 2017 and ending in April 2022
    d. Google trends data was selected to see the search behavior of individuals in that state for the following five (5) search queries. Queries were selected based on their relation to the original "flu shot" search query.
        i. "get flu shot"
        ii. "flu shot"
        iii. "flu shot near me"
        iv. "flu vaccine"
        v. "flu shot side effects"

- CDC Influenza Vaccination Coverage for All Ages (6+ Months)
    a. Source: Direct download from CDC
    b. This dataset contains monthly vaccine distribution for all 50 states in the US.
    c. The data available starts in August of 2010. Most notably, it only includes data for months considered "flu season"; as such, summer months are not included.
    d. It consists of 11 columns ad 166,000 rows
    e. Columns of interest include:
        i. State
        ii. Month
        iii. Season (year)
        iv. Dimension (sociodemographic age where estimates were calculated or for race/ethnicity)
        v. Estimate (%) (The estimated vaccine coverage or percent of people with the vaccine)
        vi. 95% CI (%) (The confidence interval of the vaccine estimation)

- Age Distribution over Sixty-Five
    a. Source: census Age
    b. Ten (10) datasets corresponding to the ten (10) states of interest were downloaded and concatenated together
    c. The percentage of the population over 65 was collected by dividing the "AGE65PLUSE_TOT_" column by the "POPESTIMATE" column.
    d. Data was collected only yearly or semi-yearly
    e. Data only went until 7/1/2022. Data projections for 7/1/2021 were manually collected from each state's respective quick facts page (*U.S. Census Bureau QuickFacts: Pennsylvania*).

- Median Household Income
    a. Source: direct download from census Income
    b. This dataset contains the annual median household income for all 50 states in the US from 2017 to 2020.
    c. The data for ten (10) states of interest was extracted from this dataset and concatenated together.

- Education Attainment
    a. Source: manually collected from census Education
    b. The percentage of people who are 25 and older and have a bachelor's degree or higher for each of the ten(10) states from 2017 to 2020 were manually collected from Census table data.

- Racial Composition:
    a. Source: directly downloaded from census Race
    b. The population by race from 2010 to 2019 for each state was downloaded.
    c. The percentage of each race was calculated by dividing the number of people of each race by the total population.
    d. Racial compositions for 2021 for each state were extracted manually from the respective data quick fact page (*U.S. Census Bureau QuickFacts: Pennsylvania*).

- Weather:
    a. Source: weather
    b. The data records the average temperature, in Fahrenheit, for each month for each state.
    c. The data is a direct download from the NOAA National Centers for Environmental Information (NCEI), but pre-processing is required to concatenate monthly data into yearly data.

- Insurance:
    a. Source: Insurance
    b. The data includes the insurance status for each state, represented by a percentage.
    c. The insurance status is categorized as Employer, Non-group, Medicaid, Medicare, Military, and Uninsured.
    d. The data is a direct download from the American Community Survey, 1-Year Estimates.

- Political Affiliation:
    a. Source: Political Affiliation
    b. The data capture the 2016 presidential election result, which represents the political affiliations of each state.
    c. The data is a direct download from Politico.

Issues:

One issue we have had during data acquisition is not having all of our data in the same time frame. Interpolation will be used to rectify this discrepancy.

In this project, we target to collect data in the time range of five years, from 2017 to 2021. However, some of the demographic data are missing for 2020 or 2021. Here is the list of the missing data:

- Median Household Income: missing data for 2021
- Education Attainment: missing data for 2021
- Racial Composition: missing data for 2020

Another issue that we had during data acquisition was that the desired demographic data was not on the same table; therefore, we needed to collect the data from several different tables on the Census website. For example, Education Attainment was manually collected and generated for each state and each year. If we decide to add more states to the project, we will need to collect the data manually once again.

Data Pre-Processing

The biggest hurdle for pre-processing is going to be matching the demographic data to monthly intervals. Because census data is only calculated once or twice a year, we will need to repeatedly use the same yearly measure or use interpolation. Interpolation is "a statistical method by which related known values are used to estimate an unknown price or potential yield of a security" (Kenton, 2020). In other words, we will extend the values we know into our monthly time frames

As mentioned in the previous part, some of the annual demographic data for 2020 or 2021 were not available. We considered several methods to solve this problem, such as using the average or using the value from the previous year. However, most of the demographic data increases slightly over a year. We plan to use a regression model to fill in the missing data to ensure having an appropriate estimation.

An additional pre-processing step will be for the google trends data. Some of the data points are "<1". They are distinct from zero and "1" values. These values will be replaced with 0.5 since there is no way of us knowing what the exact value for "<1" is.

Further, we will identify outliers and noisy data using clustering algorithms and perform correlation analysis and feature selection for data pre-processing. Feature selection will increase the predictive power of the machine learning algorithms by selecting the most critical variables and eliminating redundant and irrelevant features.

Appendix

Dataset:
https://drive.google.com/drive/folders/1YamYPb2NhafE0DF4bb1OH_lJI42Ln9Yu?usp=sharing

Linear Regression to estimate missing data Pseudocode

	A	B	C	D
1		Year	Percent Bachelor's degree or higher	
2	Massachusetts	2017	43.4	
3	Massachusetts	2018	44.5	
4	Massachusetts	2019	45	
5	Massachusetts	2020	44.5	
6	Massachusetts	2021		
7	Connecticut	2017	38.7	
8	Connecticut	2018	39.6	
9	Connecticut	2019	39.8	
10	Connecticut	2020	40	
11	Connecticut	2021		

```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()

testdf = df[df[‘Percent Bacelor’s degree or higher’]].isnull() == True]

traindf = df[df[‘Percent Bacelor’s degree or higher’]].isnull() == False]

y = traindf[‘[‘Percent Bacelor’s degree or higher’]

traindf.drop("Percent Bacelor’s degree or higher ",axis=1,inplace=True)

lr.fit(traindf,y)

testdf.drop("Percent Bacelor’s degree or higher ",axis=1,inplace=True)

pred = lr.predict(testdf)

testdf[' Percent Bacelor’s degree or higher ']= pred
```

Sources:

Abbas, K. M., Kang, G. J., Chen, D., Werre, S. R., & Marathe, A. (2018). Demographics, perceptions, and socioeconomic factors affecting influenza vaccination among adults in the United States. *PeerJ*, *6*, e5171. https://doi.org/10.7717/peerj.5171

Enten, H. (2021). *Flu shots uptake is now partisan. It didn't use to be.* | *CNN Politics*. Retrieved May 8, 2022, from https://www.cnn.com/2021/11/14/politics/flu-partisan-divide-analysis/index.html

Kenton, W. (2020, October 17). *Interpolation*. Investopedia. https://www.investopedia.com/terms/i/interpolation.asp

Larson, J. (2021). *Flu Season: When It Is and How to Prepare*. Retrieved May 8, 2022, from https://www.insider.com/flu-season

U.S. Census Bureau QuickFacts: Pennsylvania. (n.d.). Retrieved May 8, 2022, from https://www.census.gov/quickfacts/PA