Predicting Vaccine Interest Across the U.S.

Springboard Data Science Career Track
Capstone 1 Final Report
by Chantel Clark

The Problem

In the past, vaccinations have helped to control worldwide pandemics; they have saved millions of lives globally by eradicating diseases such as smallpox, polio and diphtheria (Vanderslott et al. 2019). Since the development of the vaccinations for measles, mumps, and rubella, the death rate has been reduced by 100%. Yet, there are still many children who are not vaccinated in the U.S., and this causes communities to become vulnerable to infectious diseases. In 2019 the U.S. experienced the worst outbreak of measles across the U.S. since 1992, and this was largely due to low vaccination rates in children and ultra-orthodox religious communities (Cai et al. 2019).

There has recently been a slow but steady increase in the number of children who are not vaccinated. From 2011 to 2015, the Morbidity and Mortality Weekly Report (Hill et al. 2017) reported an increase from 0.9% to 1.3% of the 2-year old cohorts were not vaccinated. While the percent increase seems small, the change in the number of children was large - the 0.4% increase means that 18,000 more 2-year olds were not vaccinated in 2015 when compared to 2011, and a total of 47,700 children in the 2-year old cohort (those born in 2013) received no vaccination in 2015. Without vaccination, children are at risk for life threatening diseases that are actually preventable.

It is difficult to obtain rates of vaccination for children for the more recent years following 2017, so federally funded and non-profit organizations such as the Vaccines for Children Program or Vaccinate Your Family whose mission is to raise awareness about vaccinations could use more frequently collected data and information to identify locations that need vaccine education the most. Google Trends API can track the Google search interest in vaccines from 2004 up to the present day in various locations across the world and could potentially use this data to decide where to allocate resources for vaccine education.

There may be many reasons for looking up vaccines on Google such as preparation for travel abroad, an attempt to gain more information about specific vaccinations for children, pets, research, etc. The list is endless, but there can still be value in finding how interest in vaccines is changing in the U.S. over time, and how interest varies by state. While Google searches can provide great resources and general information, it does not always provide credible information and many times will confuse people with conflicting information across various websites. In the case of education about vaccines, it should be provided by medical professionals.

The goal of this project is to help non-profit organizations and outreach coordinators to better predict which U.S. states have the highest need for vaccine education. Extremely high interest could be indicative of an epidemic, or uncertainty about whether or not to vaccinate a child. Low interest could be a result of lower awareness or education of the benefits of vaccination. In areas of high or low interest, surveys could be sent out to see why interest is high or low, and to also gauge interest in professional vaccine education for the community.

Questions

1. *Are there any patterns in the regions which contain high or low outliers?*
2. *How is interest in vaccines changing over time and space?*
3. *Are the interest rates across the U.S. states correlated?*

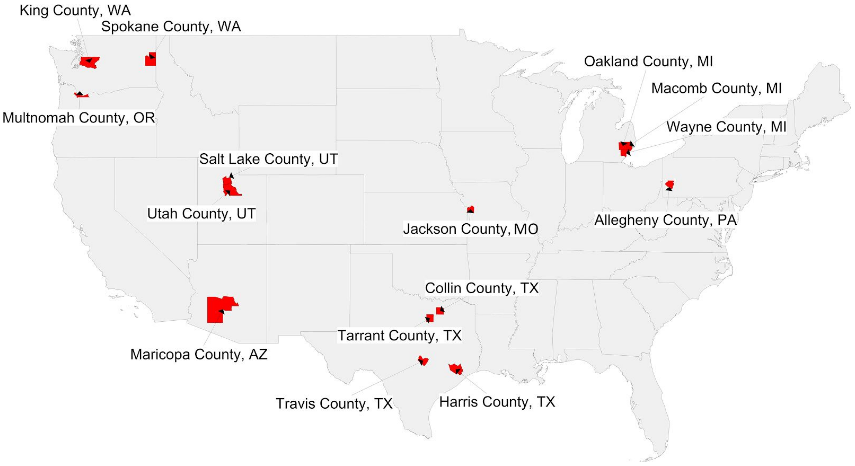
Dataset

The ‘Health searches by U.S. Metropolitan Area, 2004-2017’ dataset (Google News Lab, 2018) from Kaggle was downloaded as a CSV file and there were no missing values and did not require very much cleaning. The dataset contains ‘interest scores’ for 210 metropolitan U.S. cities in years 2004 through 2017, with a range of 0-100. Scores are representative of the fraction of all Google searches at a particular city, so a score of 100 represents very high interest, and a score of 50 means that half of all Google searches were query related. A score of 0 indicates that there was not enough data. Caution must be used in interpreting this data; it is not possible to obtain the total count of searches because the scores are proportional. This means that a larger city with half of the queries related to vaccines would receive a score lower than a smaller city where ¾ of queries were related to vaccines.

The health queries included in the dataset are cancer, cardiovascular, stroke, depression, obesity, diabetes, diarrhea, rehab, and vaccine, but only the scores for vaccine were used for this study. For potential future correlational studies with the Centers for Disease Control (CDC), the U.S. region of Northeast, South, West, and Midwest was added to the Google Health Search data set. A new column was added for state abbreviations, which were extracted from the ‘dma’ column of Google Health Search dataset, and these were then mapped to the corresponding U.S. region in another new column.

Limitations of the Data

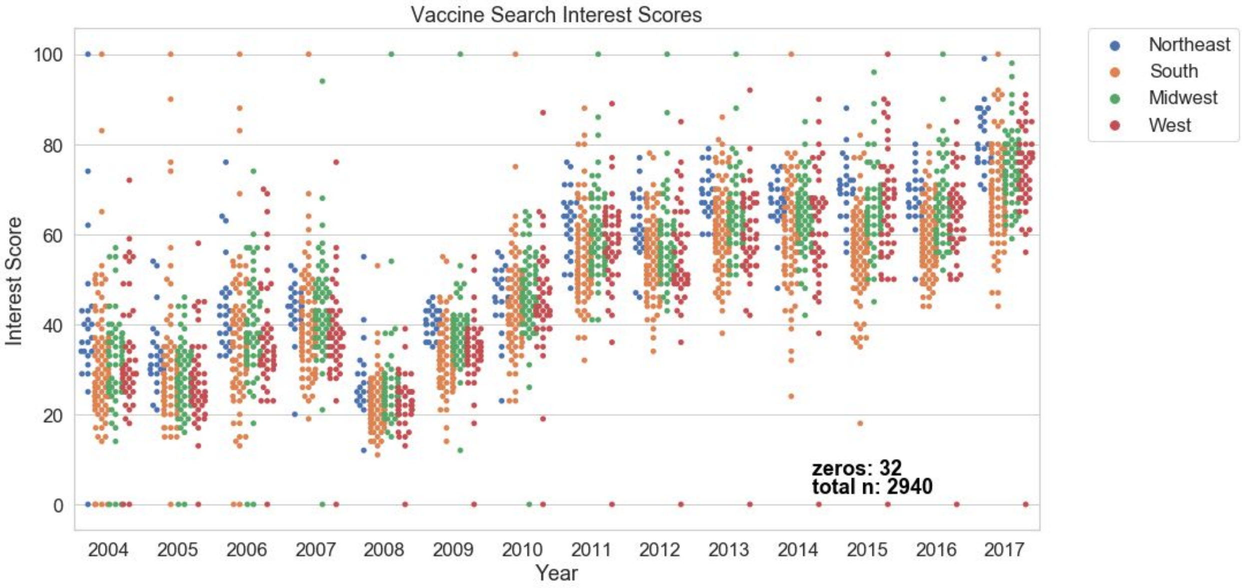
The data contains scores from metropolitan cities (MSA’s), and while children from both MSA and non-MSA’s are unvaccinated, a study in 2018 (Olive et al. 2018) has found that the largest percentage of kindergartners who obtain non-medical exemptions for vaccination come from rural areas with populations less than 50,000.



The figure above is from PLOS Medicine (Olive et al. 2018), and shows the counties within MSA's with more than 400 kindergarteners with non-medical exemptions from 2016-2017. This study could be improved if a non-MSA dataset from Google Trends were available.

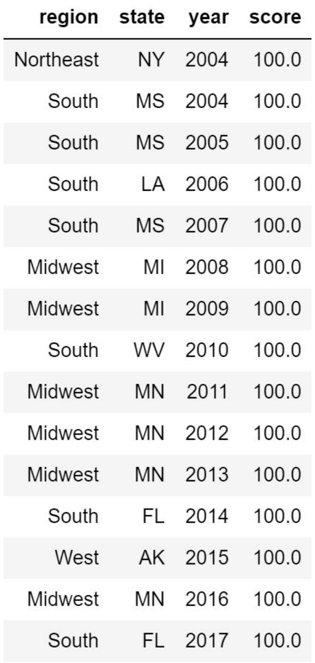
Another limitation of this study is that it was not possible to determine score fluctuations between years because the dataset contained yearly data. Monthly and weekly data is available from Google Trends API, and the model created would have better accuracy if monthly rather than yearly data was used. This could be addressed in future projects by obtaining data from the Google Trends API instead of a static dataset.

Q1) Are there any patterns in the regions which contain high or low outliers?



Within the vaccine subset of data, there were only 32 zero scores (1% of the subset). Most of these zero scores have come from the West, especially since 2007. Because the zeros comprise a small percentage of the data and do not represent 'no interest' but rather a lack of information, zeros were omitted from the dataset before computing the average state scores.

The swarm plot above allows us to see the trend of interest scores over time, and by region. There is an increase in interest from 2009 through 2017. The majority of high interest scores (100) are from the Midwest and South regions, while less are from the West, and the least from the Northeast. Prior to 2008, high interest scores were largely dominated by the Southern region.

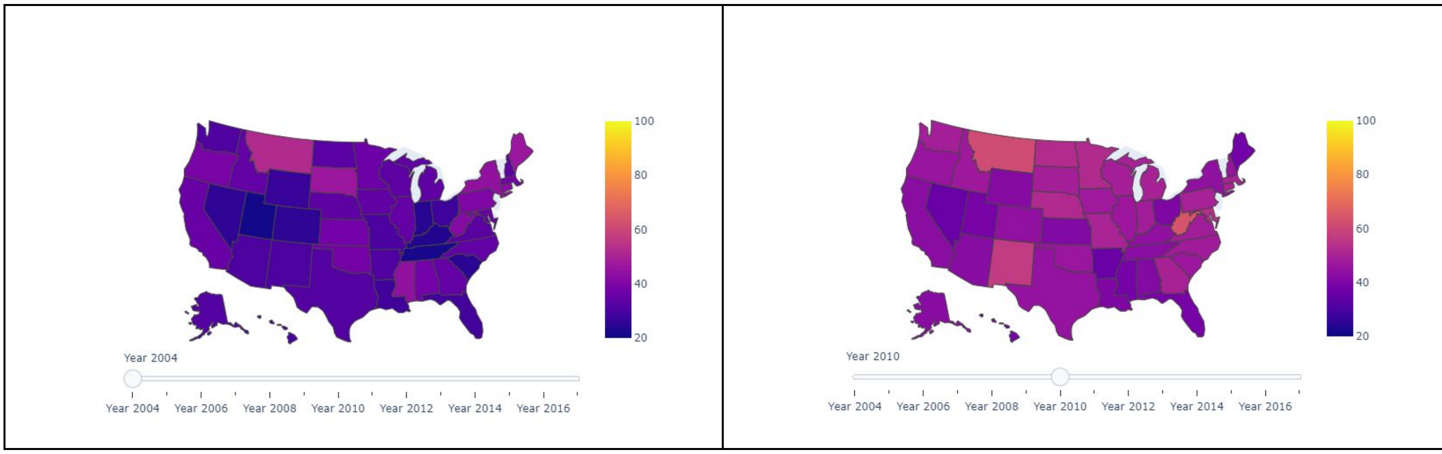


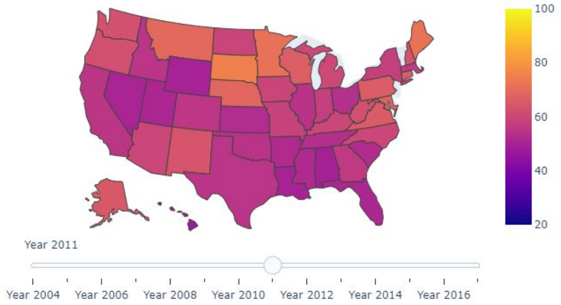
region	state	year	score
Northeast	NY	2004	100.0
South	MS	2004	100.0
South	MS	2005	100.0
South	LA	2006	100.0
South	MS	2007	100.0
Midwest	MI	2008	100.0
Midwest	MI	2009	100.0
South	WV	2010	100.0
Midwest	MN	2011	100.0
Midwest	MN	2012	100.0
Midwest	MN	2013	100.0
South	FL	2014	100.0
West	AK	2015	100.0
Midwest	MN	2016	100.0
South	FL	2017	100.0

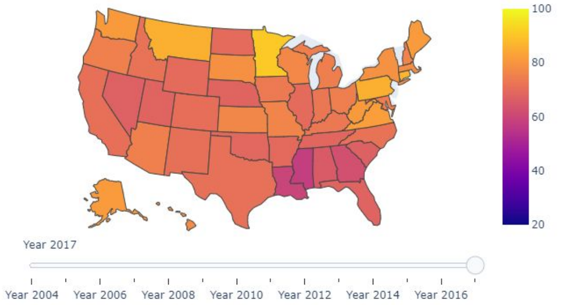
The table above shows which states contain the MSA’s with unusually high scores. Mississippi and Louisiana had high scores from 2004 through 2007, but in the most recent years, high interest scores came from the cities within Minnesota, Florida, and Alaska.

Q2) How is interest in vaccines changing over time and space?

Map of Search Interest by Year and State

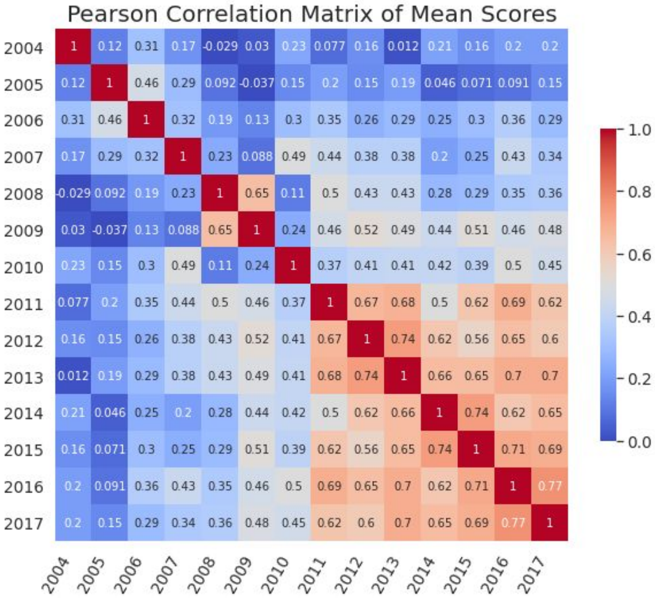






To visualize how interest in vaccines are changing over time and space, the average state score for each year was obtained by grouping the data by state and year. The maps above show the average score for various years.

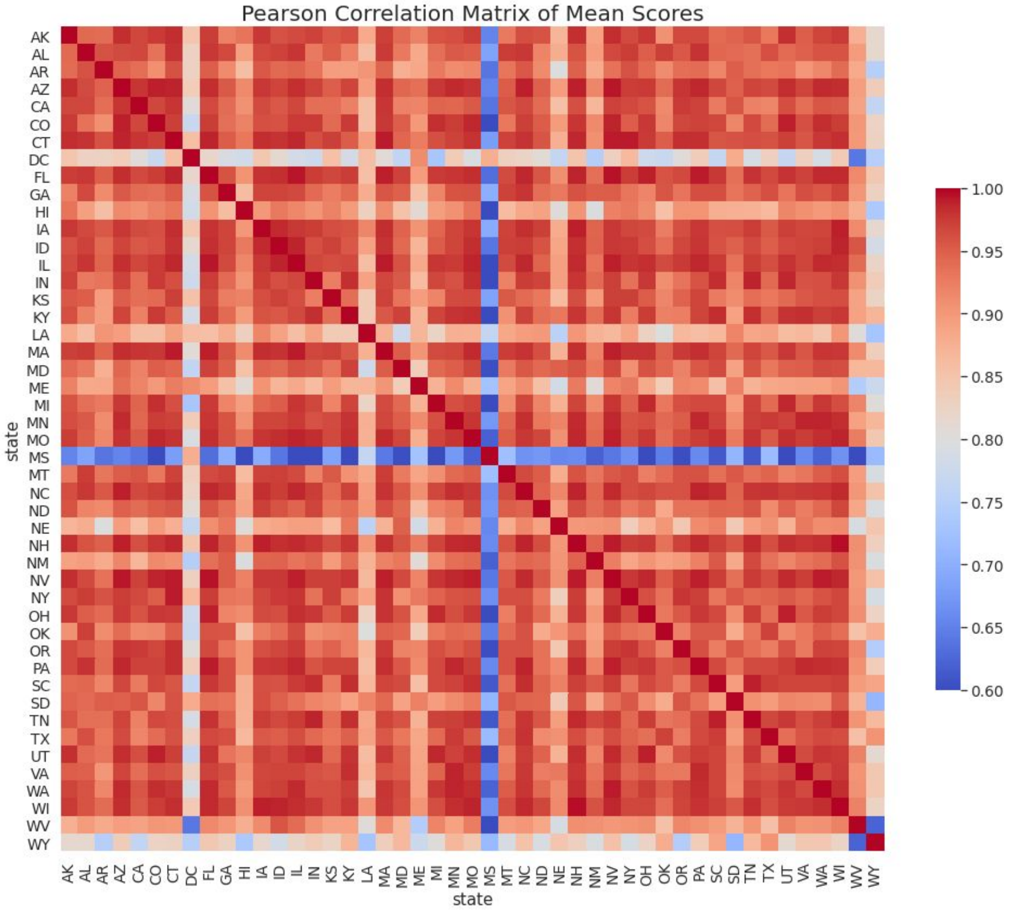
As seen in the swarm plot, interest in vaccines has generally increased over the 2004-2017 time frame. Above are a few maps of the various years, and the brightening of colors represents an increase in interest. These maps can be displayed online with an interactive slider to view each year, and also contains a hover feature that displays the name of the state and average interest score. These maps allow for visualization of how the scores for individual states are changing over time. While there has been an overall increase in interest across the U.S., some states such as Wisconsin and Alaska have drastically increased interest. These maps also allow for identification of which states have the lowest average interest score. In 2017, Mississippi and Louisiana have the lowest interest scores.


Pearson Correlation Matrix of Mean Scores

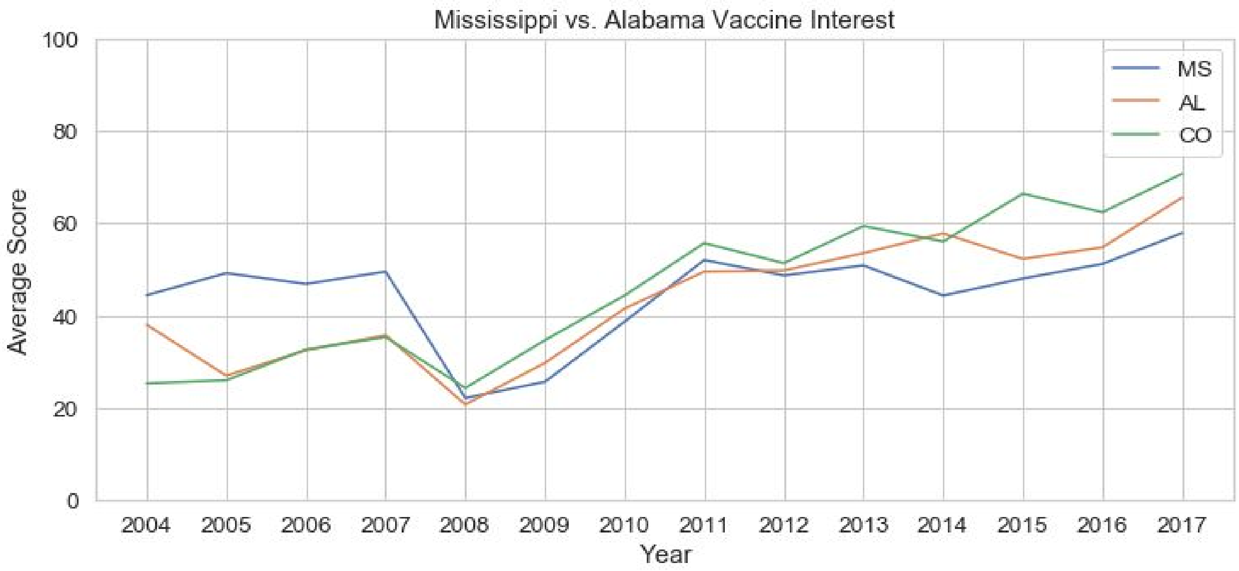
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2004	1	0.12	0.31	0.17	-0.029	0.03	0.23	0.077	0.16	0.012	0.21	0.16	0.2	0.2
2005	0.12	1	0.46	0.29	0.092	-0.037	0.15	0.2	0.15	0.19	0.046	0.071	0.091	0.15
2006	0.31	0.46	1	0.32	0.19	0.13	0.3	0.35	0.26	0.29	0.25	0.3	0.36	0.29
2007	0.17	0.29	0.32	1	0.23	0.088	0.49	0.44	0.38	0.38	0.2	0.25	0.43	0.34
2008	-0.029	0.092	0.19	0.23	1	0.65	0.11	0.5	0.43	0.43	0.28	0.29	0.35	0.36
2009	0.03	-0.037	0.13	0.088	0.65	1	0.24	0.46	0.52	0.49	0.44	0.51	0.46	0.48
2010	0.23	0.15	0.3	0.49	0.11	0.24	1	0.37	0.41	0.41	0.42	0.39	0.5	0.45
2011	0.077	0.2	0.35	0.44	0.5	0.46	0.37	1	0.67	0.68	0.5	0.62	0.69	0.62
2012	0.16	0.15	0.26	0.38	0.43	0.52	0.41	0.67	1	0.74	0.62	0.56	0.65	0.6
2013	0.012	0.19	0.29	0.38	0.43	0.49	0.41	0.68	0.74	1	0.66	0.65	0.7	0.7
2014	0.21	0.046	0.25	0.2	0.28	0.44	0.42	0.5	0.62	0.66	1	0.74	0.62	0.65
2015	0.16	0.071	0.3	0.25	0.29	0.51	0.39	0.62	0.56	0.65	0.74	1	0.71	0.69
2016	0.2	0.091	0.36	0.43	0.35	0.46	0.5	0.69	0.65	0.7	0.62	0.71	1	0.77
2017	0.2	0.15	0.29	0.34	0.36	0.48	0.45	0.62	0.6	0.7	0.65	0.69	0.77	1

Because data is a time series, consecutive years will have a tendency to be correlated. For example, data in 2017 is likely to be related to the data in 2016. In the correlation matrix above, we can see that there is not much correlation between the years from 2004 through 2007. Scores in 2008 were moderately correlated with 2009, while scores from 2011 through 2017 showed the highest correlation.

Q3) Are the interest rates across U.S. states correlated?



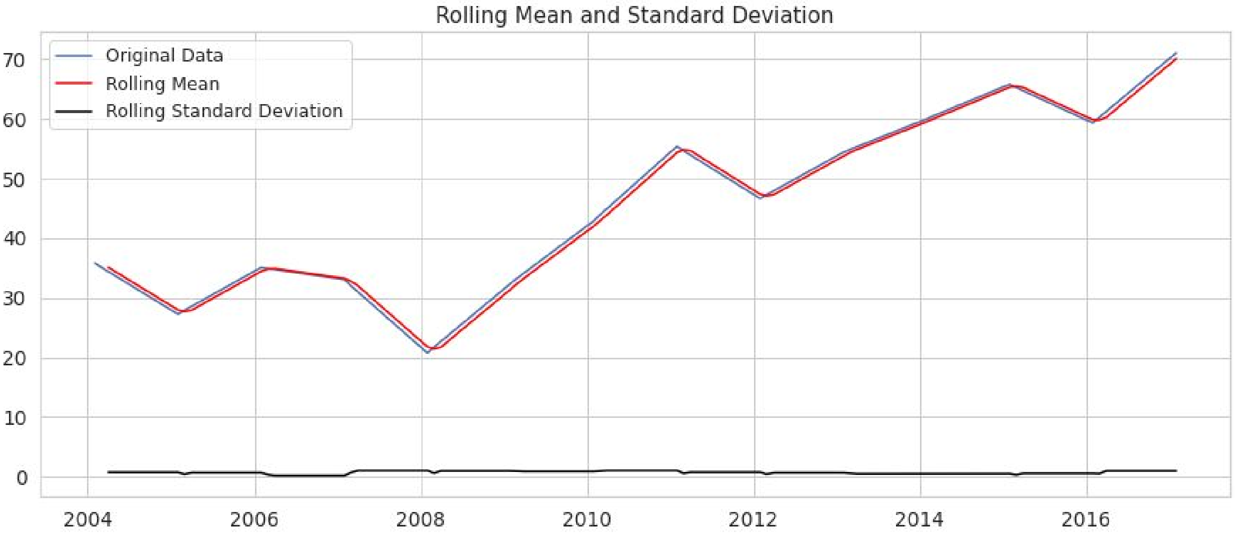
Most states correlated highly with the other state, but Mississippi had the lowest correlation scores around 0.60. To see how different the scores for Mississippi are from the others, a line graph of the scores over time were compared with the scores from Alabama and Colorado.



It is apparent from the plot above that Alabama and Colorado are more similar than Mississippi in score value and general trend, but the scores from Mississippi do not seem very abnormal. All three states have a low score in 2008 that increases through 2017. This demonstrates that Mississippi's lower correlation score of 0.60 still has moderate correlation with the other states.

Statistical Analysis for Model

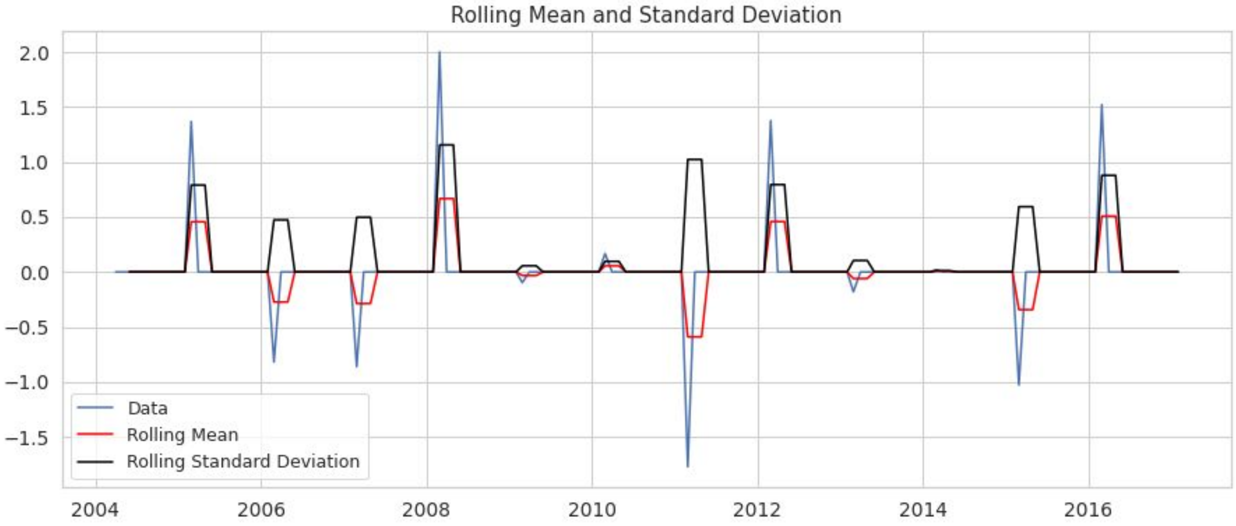
A Dicky-Fuller test was used to determine whether or not the data has any trends (is the data stationary). The null hypothesis is that the time series has a unit root (is non-stationary), or some type of dependent structure. The alternative hypothesis is that the time series does not have a unit root (is stationary), and has no trend.



```
Test Statistic           -0.389136
p-value                   0.911897
n_lags                   13.000000
n_observations          143.000000
Critical Value (1%)      -3.476927
Critical Value (5%)      -2.881973
Critical Value (10%)     -2.577665
```

The p-value is 0.91 for the original data, which means that assuming the condition of the null-hypothesis (that the data has a trend) there is about a 91% chance that we would obtain data as seen here. This means that we fail to reject the null hypothesis and can conclude that there is in fact a trend in the data.

When creating a predictive model, it is important to have trends removed, therefore data was differenced. Differencing the data finds the difference between each successive pair of scores. For example, if data in a time series were [5,10,8,11] the differenced data set would be [5,-2,3]. With one difference the Dicky-Fuller test resulted in a p-value of 0.14, so this removed most of the trend, but not all.



```
Test Statistic          -6.089455e+00
p-value                  1.044985e-07
n_lags                   1.100000e+01
n_observations           1.430000e+02
Critical Value (1%)     -3.476927e+00
Critical Value (5%)     -2.881973e+00
Critical Value (10%)    -2.577665e+00
```

The result of differencing twice is shown above, where a p-value of 1.04e-7 is obtained. Because the p-value is very close to zero, we can reject the null hypothesis and conclude that the data is stationary. The Dicky-Fuller test confirms that we must use an autoregressive model to account for the non-stationarity of the data.

Machine Learning Model

An Autoregressive Integrated Moving Average (ARIMA) model was used to model the mean search interest score for vaccines. The ARIMA model has parameters (*p,d,q*), where *p* is the lag value for autoregression, *d* is the difference order, and *q* is the size of the moving average window. The data given were annual scores, and there were a total of 14 scores per metropolitan city for years 2004-2017. Data was smoothed by upsampling time indices by month, and values were interpolated linearly. To find starting values of the parameters, the autocorrelation functions (ACF) and partial autocorrelation functions (PACF) were plotted for the original data, once differenced data, and twice differenced data.

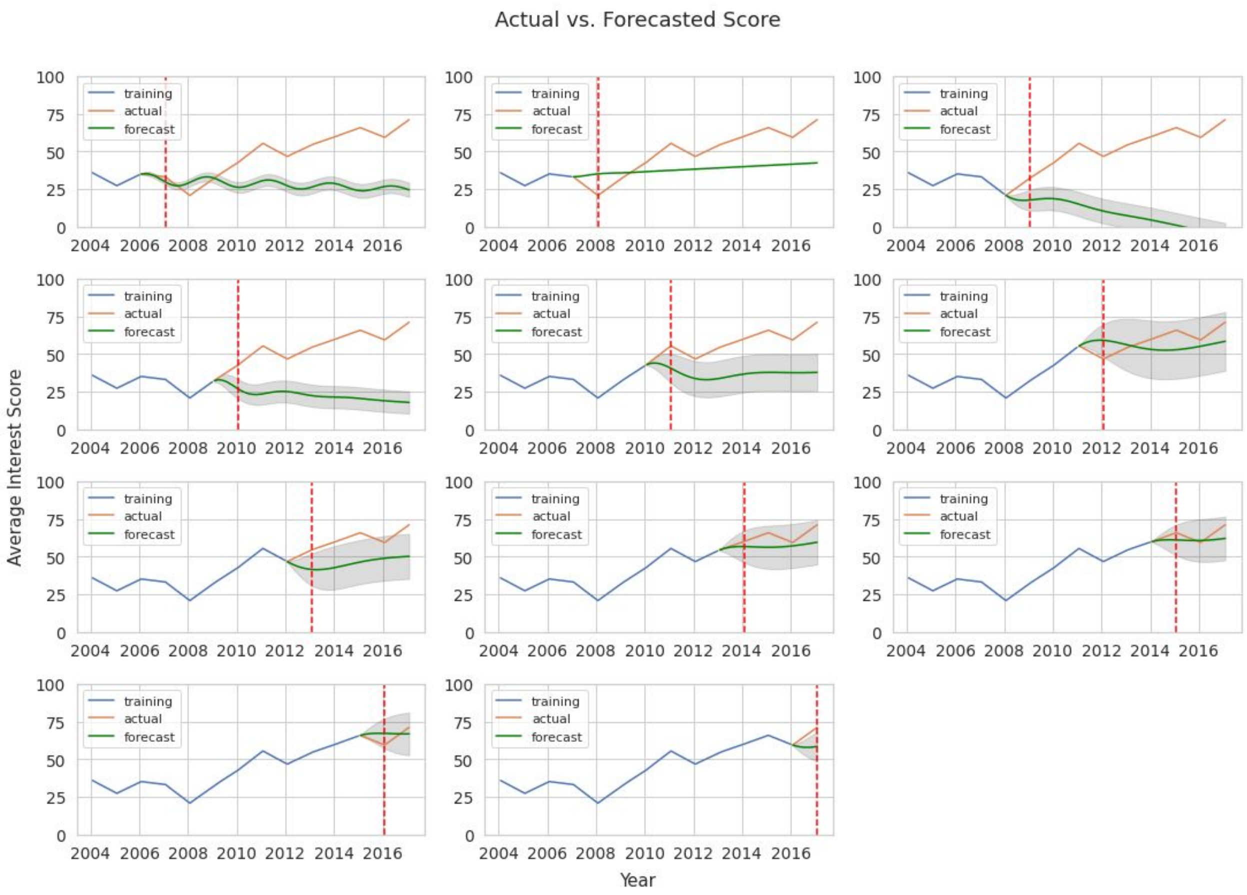
The figure above shows the mean interest scores from California, once and twice differenced data, along with an associated ACF plot. There is clearly an upward trend in the original data, a small upward trend after differencing once, and no trend after differencing twice. The first difference looks like it accounts for autocorrelation well because after lag 5, no other lags are outside of the significance range (shaded area). The plots above indicate that *d*=1 and *q*=5 might be a good place to start.

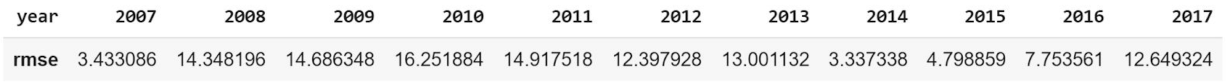
From the PACF plots above, we can see that the first lag is outside of significance bounds, therefore a good first step is to try using 1 lag for parameter *p*. With order (1,1,5), the AIC score is 94.671, and we will compare this value with other orders of the ARIMA model. Minimizing the AIC score is indicative of a better fit model.

A loop was created to determine which combinations of *p*,*d*, and *q* optimized the model - in essence, which model gave the lowest Akaike Information Criterion (AIC) score. The best AIC score was from using ARIMA order of (5,1,4), followed by (5,1,3). The scores were 81.366 and 84.318 respectively. Both of these scores are significantly better than the initial model with order (1,1,5).

Looking at the p-values of coefficients, the first model with order (5,1,4) has p-values of zero except for 2 p-values greater than 0.8 for an autoregressive and moving average coefficient. The order (5,1,4) model also resulted in p-values of zero except for 2 p-values between 0.1 and 0.45. Both models have a couple of coefficients that are not playing a significant part in the model, but most of the coefficients look good with p-values near zero.

The models were also evaluated by calculating the root mean square error (RMSE) for 1-year forecasts. This was done by splitting the data into two contiguous sets. For example, the first trial used training data from 2004-2006, and the RMSE for 2007 was calculated by finding the difference between the true score and the predicted score for 2007. This process was repeated for predictions of years 2007 through 2017, with the size of training data increasing with time because all data prior to the testing year was used as the training data set.

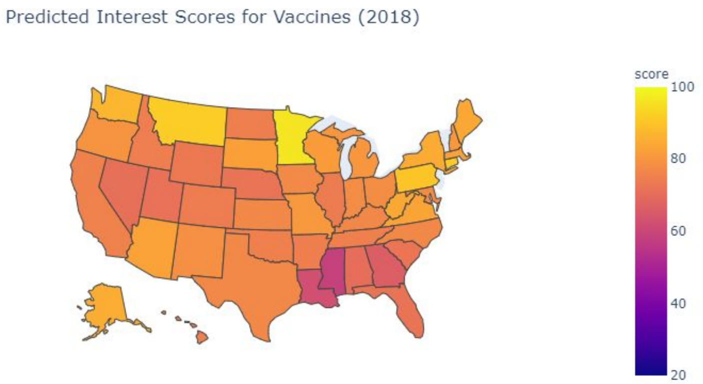




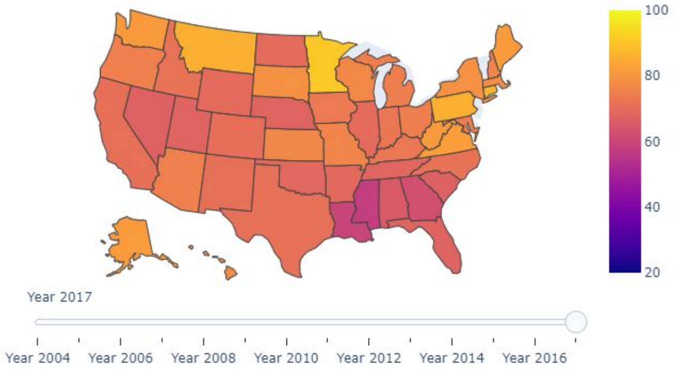
year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
rmse	3.433086	14.348196	14.686348	16.251884	14.917518	12.397928	13.001132	3.337338	4.798859	7.753561	12.649324

Using the ARIMA model with order (5,1,4) for California scores, the RMSE for each year was calculated (table above), and has a mean of about 10.7. This means that on average, the 1 year out prediction based off of all previous data is about 11 points off from the true value.

The process of tuning parameters for the ARIMA model was used for every state in the data set to create a map of predicted scores in 2018.



For comparison, the 2017 search interest scores are shown below.



Summary

Q1) Are there any patterns in the regions which contain high or low outliers?

From 2004-2017, the South and Midwest regions have had the highest levels of interest in vaccines. In the most recent years, cities in Minnesota, Florida, and Alaska have had the highest interest in vaccines. Places that received zero scores (not enough information) largely came from the West since 2007.

Q2) How is interest in vaccines changing over time and space?

Interest across the U.S. has grown tremendously from 2008 through 2017, with some states showing more changes than others. Large increases in the average state interest scores were seen in Alaska and Wisconsin, with less change in Louisiana and Mississippi.

Q3) Are the interest rates across the U.S. states correlated?

Yes, the U.S. states have moderate to high correlation, and this can be seen in the state to state correlation matrix and the overall growing interest scores over time. There has been increasing interest in vaccination across the entire U.S.

It would be interesting to compare these predictions to the Google Trends API data from 2018, and also to apply this model with a finer resolution over time (using monthly rather than yearly data) and space (using city versus averaged state data). This project has the potential to be turned into an app with the Google Trends API. It could be used by health organizations interested in increasing vaccination rates to visualize real-time interest levels and predictions across the U.S. Health organizations could determine which U.S. states have the highest (or lowest) interest, or rapidly increasing (or decreasing) interest in vaccination. This could help these healthcare organizations to decide where resources should be allocated for maximum efficacy. Equipped with more information about how people in the U.S. are thinking about vaccination, perhaps rates of vaccination of children can be improved.

References

Cai W., Lu D., and Reinhard S. (2019 June 3). *Largest U.S. Measles Outbreak in 25 Years Surpasses 980 Cases.* The New York Times Company. Retrieved on 2020 March 30 from https://www.nytimes.com/interactive/2019/health/measles-outbreak.html

Google News Lab (2018). Health searches by US Metropolitan Area, 2005-2017. Retrieved on 2020 March 30 from https://www.kaggle.com/GoogleNewsLab/health-searches-us-county

Hill H.A., Elam-Evans L.D., Yankey D., Singleton J.A., Kang Y. (2017) Vaccination Coverage Among Children Aged 19–35 Months. *MMWR Morb Mortal Wkly Rep* 2018;67:1123–1128. Retrieved on 2020 March 30 from DOI: http://dx.doi.org/10.15585/mmwr.mm6740a4

Olive JK, Hotez PJ, Damania A, Nolan MS (2018 June 12). The state of the antivaccine movement in the United States: A focused examination of nonmedical exemptions in states and counties. *PLOS Medicine 15*(7): e1002616. Retrieved on 2020 March 30 from https://doi.org/10.1371/journal.pmed.1002616

Sun, Lena H (2018 June 20). *Kids in these U.S. hot spots at higher risk because parents opt out of vaccinations*. The Washington Post. Retrieved on 2020 March 30 from https://www.washingtonpost.com/news/to-your-health/wp/2018/06/12/kids-in-these-u-s-hotspots-at-higher-risk-because-parents-opt-out-of-vaccinations/?noredirect=on

Vanderslott S, Dadonaite B, and Roser M (2019 July). *Vaccination*. OurWorldInData.org. Retrieved on 2020 March 30 from , https://ourworldindata.org/vaccination