**Predicting Pertussis (Whooping Cough) Outbreaks in California Schools**

Ahmed Mohamed, Byron Dickerson, Clara Comstock, Rick Djeuhon, Julia Goyco

**Proposal**

After exploring many datasets, and evaluating possible solutions and relevance to societal

need, we decided to dive into the topic of health specifically disease outbreaks. We

narrowed down the general topic and chose one dealing with California immunization

rates, in which we found a large dataset dealing with California Pertussis cases and rates from

2010 to 2014. The dataset includes number of students polled, school name, school county, the

number of students with a personal belief exemption from vaccines, the number of students with a medical exemption from vaccines, the number of students that had received specific types of vaccination, and if the school was private or public. Other data sets attached with the problem include the geographical coordinates of each school and the yearly rate of pertussis outbreaks by county. These datasets are provided to Kaggle by the California Department of Public Health (CDPH). Upon inspection of the dataset, we found it interesting to examine what role vaccination rates in every county played in infant Pertussis cases, as well as how the reported exemption for a given school affected the outbreak.

Our question is: can we predict which counties in California are at a higher risk of

Pertussis outbreaks based on previous years’ vaccination records? This question would be

of societal need as the Pertussis cases and rates over the years lead to infant hospitalization and,

unfortunately, some deaths. In addition, if we can build a predictive model that uses previous

years to predict the upcoming year’s risk, then early intervention and an increase in vaccination rates can help combat the issue and reduce outbreak cases. Furthermore, the model will be tested and validated on the 2014 year to help further improve the model and set possible modifications for future improvements of the model.

**Introduction**

Pertussis is an illness that is contagious and affects all ages with the highest risk of harm/death being ages less than five, as concluded by BMC Research Notes. Despite the fact that the highest factor contributing to pertussis development deals with infected persons in the family/household, the highest protective mechanism against developing pertussis remains through vaccination (Alamaw et al, 2017). Thus, the uniqueness of this data science project is to forecast and predict pertussis high risk outbreaks based on data dealing with student vaccinations.

In an attempt to conquer the issue described, the team decided to break down the problem into three categories which are data mining, data visualization, and machine learning. The first category of the project dealing with data mining would help understand the data through geographical depiction. From the geographical representation the data can be prepared and cleaned accordingly. The second category focuses on how the data can be statistically represented in a manner that would help tackle the questions: do public schools or private schools have higher vaccination rates, are vaccination rates increasing or decreasing, which county has the highest outbreak rate, which variables mainly affect the number of pertussis cases, and which variables could be thrown out? The last category of the project deals with the machine learning aspect specifically the use of regression and classification to quantify and predict the risk of outbreaks per county, and thus anticipate which counties are at a higher risk.

**Available Data and Conclusions**

Many data are already available dealing with Pertussis and Pertussis vaccinations as the disease spreads easily, quickly, and affects all ages especially younger ages. Example of such data deals with number of vaccinations, number of cases, and vaccination effectiveness. Also, some conclusions are already available dealing with some of the questions to be explored such as the relationship between public and private schools and how private schools tend to have more exemptions and thus less vaccinations compared to public schools (Shute, 2013). Also, conclusions on the effectiveness of acellular pertussis vaccines compared to whole cell pertussis vaccines are available which show whole cell pertussis vaccines tend to be more effective with some an example of dropping pertussis cases from average of 771 to 21 per 100,000 (Awaidy, 2018), while acellular pertussis vaccines tend to cause less reactions to the vaccination (Korkmaz et al, 2013) and also less adverse events following immunization (Patterson et al, 2018).

**Data Cleaning**

In the data cleaning process, the first step was to drop columns that were of no use for building the predictive model, those columns were school type, school name, school code, number of students reporting personal belief exemption, and number of students reporting medical exemption. The next step in the data cleaning process involved keeping only the years that can be used for training and testing as well as dropping further columns in the student data CSV file, upon inspection of the pertussis rates CSV file, it was observed that it tracks cases and rates over the 2010 to 2014 years, and also it tracks those rates and cases for only pertussis disease. However, the student data CSV file maintains years of data from 2000 to 2015, and it also contains vaccination data for other diseases, as a result the years 2000 to 2009 as well as 2015 were dropped from the student data, and also the number of students reporting MMR (Measles/Mumps/Rubella disease) vaccination and number of students reporting polio vaccination were dropped.

The second step in the data cleaning process involved sorting and ordering the new student data CSV file. The looping structure to be used for the calculation of averages of vaccinated, unvaccinated, and cases over the 2010 to 2014 years required the data to be sorted per county similar to the ordering of the rates data CSV file (alphabetically). Also, the sorting should keep the data of 2010 to 2014 years per county in the new alphabetical county order, meaning that each county would display in the CSV file the years 2010 to 2014 before moving to the next county. Such sorting and ordering were handled by specifying the sort according to the county and the year respectively.

The third step in the data cleaning process executed the calculation of the number of unvaccinated students in every county, every school, and every year. Such calculation was simply executed by subtracting the total number of students per county for each school and each year to its respective number of students reporting DTP vaccination, the resultant was a stored and added as an extra column to the student data frame.

The fourth step in the data cleaning process revolved around retaining and then removing the data of vaccinated and unvaccinated numbers for the year 2014. As the predictive model to be built would be trained on years 2010 to 2013 then tested on 2014, the 2014 data would need to be retained for the testing process but also removed from the student data CSV file. Such operation required a simple looping structure, which was executed to the length of the total number of rows for the years in the student data CSV file, then a simple check of the current row’s year against 2014 was performed and once found, the current index was used to retrieve the vaccinated and not-vaccinated data and then those data were added to two new lists of vaccinated and unvaccinated accordingly, then a new dataframe was created with columns containing the vaccinated and unvaccinated lists which would later be used for the testing process. After such operation, the data for 2014 in student data frame were dropped, as the data would then be later further modified and used for the training process of the model. In addition, upon further inspection of the result of the lists of 2014 data, the county Alpine was observed to be missing completely for the 2014 year, as a result the county was appended with zero values for the 2014 year to prevent shape errors that would affect the testing process.

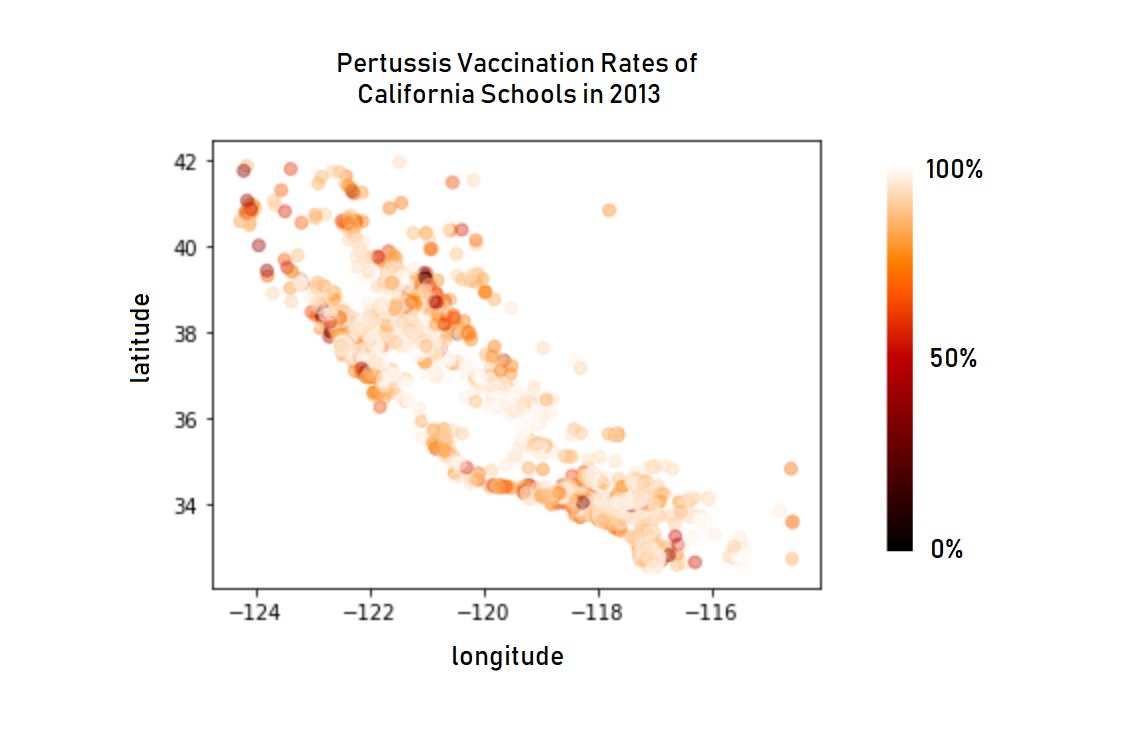
The final step in the data cleaning process involved the calculation and creation of a new dataframe that would contain averages of vaccinated, unvaccinated, and cases over the 2010 to 2013 years. As the updated student dataframe contains each county from years 2010 to 2013 before moving to the next county, a simple looping structure was performed in which a skip of four elements was performed (to take the loop county by county for each year from 2010 to 2013) and in each loop iteration, the values for vaccinated, unvaccinated, and cases were gathered for each county and the average of each of those values was calculated and stored in a list, this way the list contains averages of the data from 2010 to 2013. Finally, a new dataframe was created that held those lists of averages as columns, and that dataframe represents the data that would be used for the training process for the predictive model.

**Data Analysis**

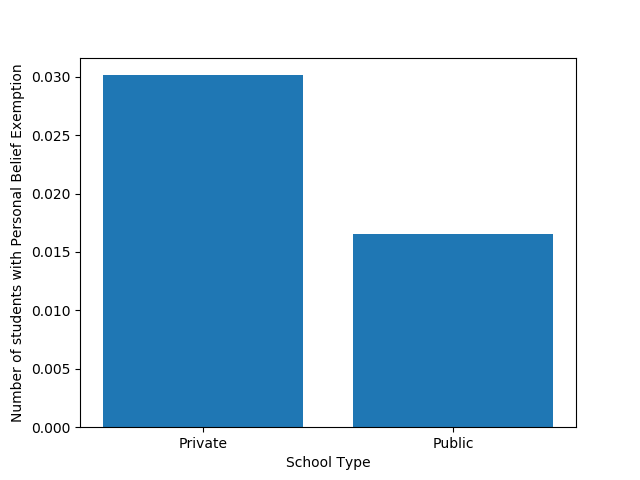
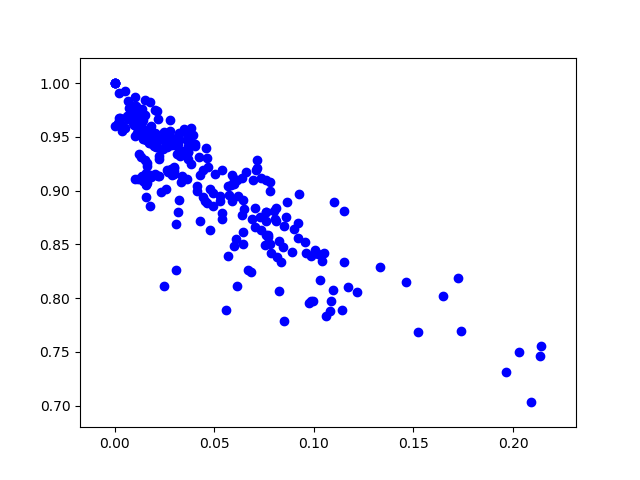
For the data analysis portion of the project we did exploratory data analysis and kMeans clustering to find trends in the data. We made plots to identify key features that affect the number of pertussis cases. Using kMeans clustering we found there was a possible relationship between the distance between schools and the rate of pertussis outbreaks. Some question we had about the data were: do public Schools or private schools have higher vaccination rates, are vaccination rates increasing or decreasing, and which county has the highest outbreak rate? This helped us decide what model to use, which variables affected the number of pertussis cases, and which variables could be thrown out.

K-means clustering of the school geographical data revealed two distinct clumps of school locations, as well as a wide smattering of locations too far from the other schools for proximity to play any sort of major role in pertussis outbreak spread. One difficulty faced was the inability to map these coordinates onto any actual roadmap of California. However, these clusters did inspire the next step for further geographic exploration.

The next major step in visually representing the data was to compute the (percentage) vaccination rates of each school provided in the student data set and map that information onto the geographical coordinates. This provided a convenient visual reference for determining which schools had lower rates, as well as their proximity to each other. A different map of this data was made for each year within our range of consideration to further assist trend recognition. These charts were made not for numerical analysis, but rather to suggest good starting off points for further analysis.



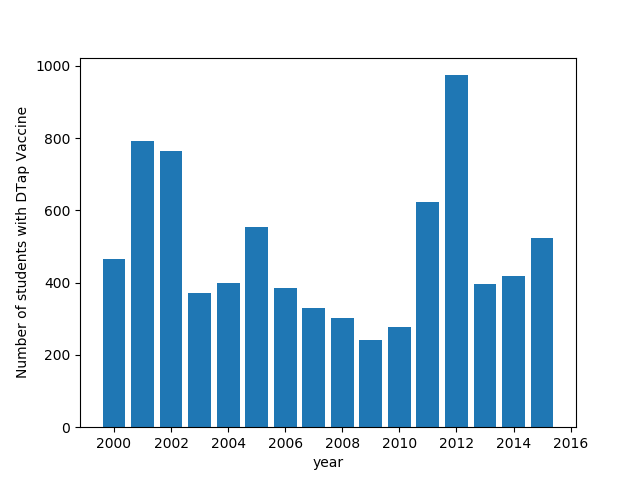
Next, for the exploratory data analysis portion of the project, the team made pairplots of the data. Pairplots enable the construction of histograms and scatter plots. The histograms show the distribution of a single variable and the scatter plots shows the relationship between two variables. From these plots, we are able to see what does have a relationship and what doesn’t. The pairplot shows a linear relationship between if students have received the DTP (Diphtheria/Tetanus/Pertussis) vaccine, MMR(Measles/Mumps/Rubella) vaccine, and the Polio vaccine. These variables most likely have a linear relationship because if a student gets one vaccine, then they are more likely to get the other vaccines. We also saw there was a negative correlation between the number of kindergarten students who got the pertussis vaccine and the number of kindergarten students who claimed to be exempt from vaccines due to personal beliefs. Figure 1 shows that if a student did not get the DTP vaccine, then they probably have a personal belief exemption.



**Figure 2:**Bar chart shows private school children claim a personal belief exemption more than public school children.

**Figure 1:**Scatter plot of number of students who got pertussis vaccine and number of students who claim personal belief exemptions between 2010-2014.

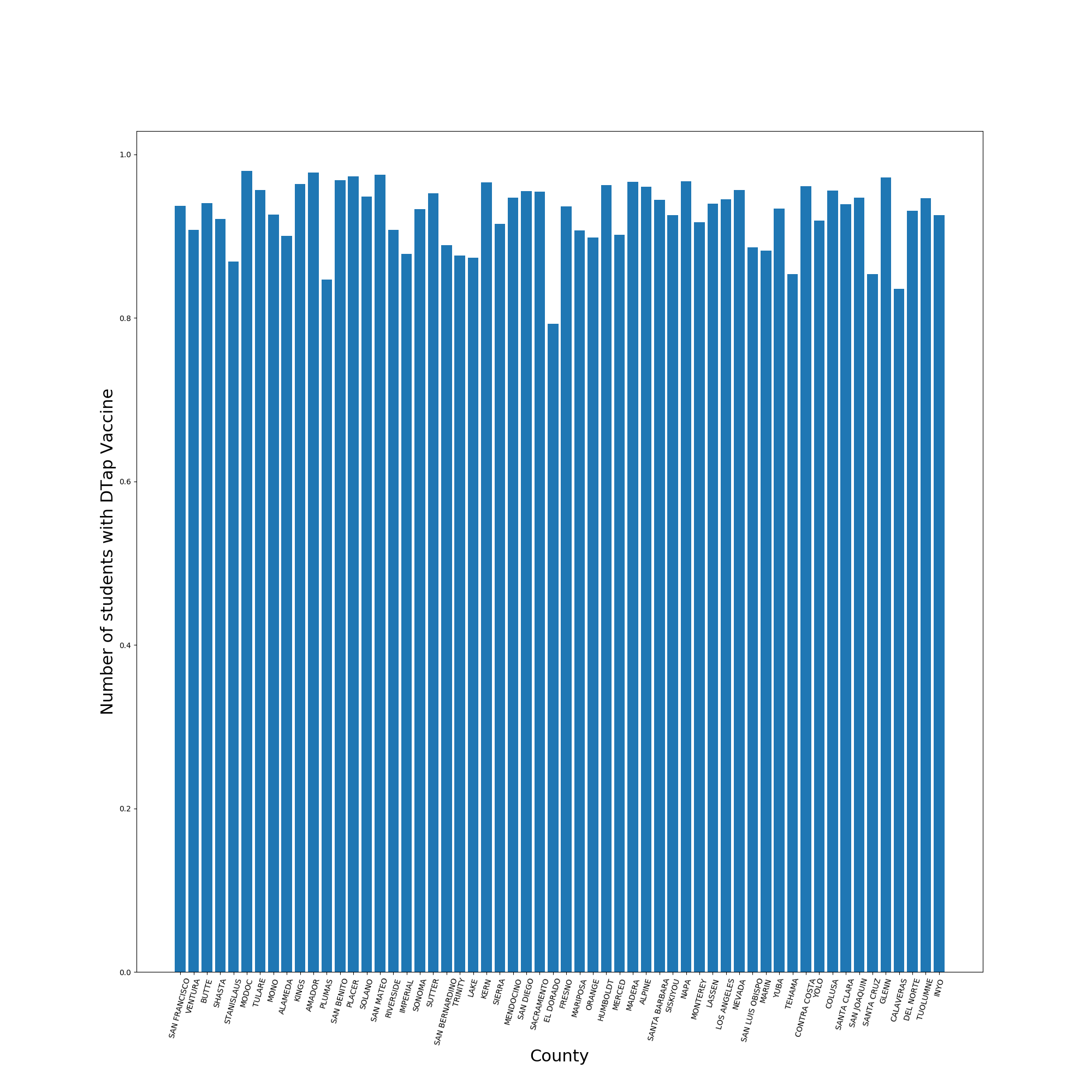
Figure 2 shows a higher rate of personal belief exemptions in private schools. This, in turn, means children in private schools are not getting vaccinated as much as public school children. This is consistent with research done by graduate students at Johns Hopkins Bloomberg School of Public Health. They found that both exemptions and clusters of pertussis cases tended to be in neighborhoods with higher levels of education and income (Shute, 2013).





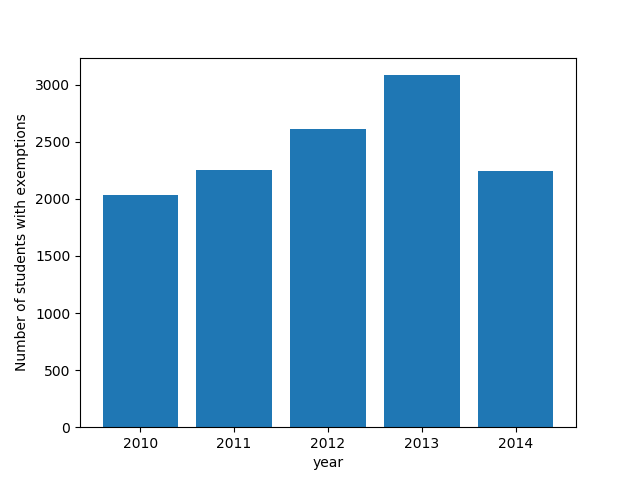
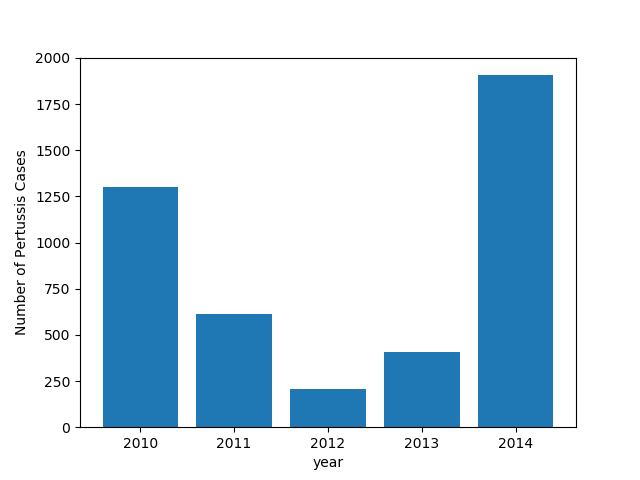
We decided to plot the number of students who got the pertussis vaccine for the years 2000-2014 to see if there was any trend. The first low point in the data occurs in the year 2000. This might be due to a paper published in 1998 in the Lancet medical journal. This paper implied a link between vaccinations and autism. In the original paper, Wakefield and 12 co-authors claimed to have investigated 12 children referred to the Royal Free Hospital and School of Medicine with chronic enterocolitis and regressive developmental disorder. The authors reported that the parents of eight of the 12 children associated their loss of acquired skills, including language, with the MMR vaccination. The authors concluded that the MMR vaccine was associated with the onset of both the gastrointestinal disease and developmental regression (Eggertson, 2010).

The paper was first investigated by the Lancet in 2004. Twelve years later the paper was retracted, but to this day many autism advocacy groups and parents continue to defend the paper’s findings. We found that the 2009-2010 time period had some of the lowest vaccination rates. This is due to parents claiming personal belief exemptions. At some schools over a third of the population was not vaccinated. Researchers have proven that the low vaccination rate was a major factor in the pertussis outbreak in 2010 (Shute, 2013).



**Figure 4:**Average Number of students who got DTaP vaccine in each county.

The counties all appear to have high pertussis vaccination rates but if the vaccinated population drops below 92-94% then herd immunity is lost. ‘Herd immunity’ refers to how, when a significant portion of a population is immune to a disease, there are not enough vulnerable people around to carry and spread it. Because the disease cannot spread to them, the vulnerable members of the population remain safe (they are “immune”). The estimated coverage needed to prevent the spread of whooping cough is 92-94% (Helft et al, 2014). Figure 4 shows that a lot of counties are at or below 90%. This creates pockets of unvaccinated children and allows the disease to spread quickly in those areas.



**Figure 5:**Total Pertussis cases for the years

2010- 2014

A pertussis outbreak is cyclical, with peaks every 3-5 years. In 2010, approximately 9,000 cases were reported, including 808 hospitalizations and 10 infant deaths, for a statewide incidence of 24.6 cases per 100,000 population (Winter et al, 2014). The last pertussis epidemic happened in 2014. A total of 9,935 cases of pertussis were reported in 2014, for a statewide incidence of 26.0 cases per 100,000 (Winter et al, 2014). The number of pertussis cases in 2014 was larger than in 2010. In 2014, there was an increase in cases of 14-16 year olds, but their hospitalization rate is low compared with that of infants. They were the first to receive acellular pertussis vaccines, used exclusively since 1997, and while acellular pertussis vaccines provide less local, systematic, and febrile reactions (Korkmaz et al, 2013), they are considered to have less-durable immunity (Lehman, 2014). The acellular vaccination was shown to only be more effective for ages less than 7 years, but not the ages that suffered in the epidemic (Plans et al, 2016).

Another factor that determines the number of pertussis cases is the effectiveness of the pertussis acellular vaccine versus the whole cell vaccine. The highest number of pertussis cases from California outbreaks in 2010 and 2014 were infants. The second largest category was the older children and adolescents, having average age of 10 (2010) and 14–16 (2014). This indicates waning immunity among older children and adolescents. All of these children would have received the acellular vaccine.

**Predictive model**

Lastly, in the task of building the predictive model to tackle the machine learning aspect of this data science project, two models were built and compared with their respective accuracy with the goal of choosing the better predictive model for the project. Both predictive models relied on machine learning techniques operating on the numerical data dealing with pertussis cases per county over the 2010 to 2014 years, and both models were executed with the intent of predicting which counties in California are at higher risk of pertussis cases outbreaks. However, the way the data was cleaned and the actual machine learning technique used for each model differed. The first model used regression to predict and produce pertussis cases based on vaccinated and unvaccinated numbers of students as pairs relative to overall number of pertussis cases, while the second model used classification to sort each county’s risk according to unvaccinated numbers of students.

Both predictive models utilized the student data CSV file, which contains a great deal of data relating to California schools per county. Such data has the school type (whether the school is public or private), county name, school name, total number of students per school, school’s code (a unique integer used to identify schools over the years), number of students reporting polio vaccination, number of students reporting MMR vaccination, number of students reporting DTP vaccination, number of students reporting personal beliefs exemption, number of students reporting medical exemption, and the years of when all of those data points were gathered (ranging from 2000 to 2014). In addition, both models used the pertussis rates CSV file, which contains pertussis cases and rates for every county in California from the years 2010 to 2014.

The first predictive model used the approach of training on years from 2010 to 2013 on the student data CSV file with only the numerical values calculated of vaccinated, unvaccinated, and pertussis cases. Those numerical values were averaged over the 2010 to 2013 years. Then, the model used a 70 to 30 training to testing split ratio to train and test on the data. Using linear regression, the model predicted value outcome of cases based on the vaccinated and unvaccinated data for the year 2014, and the resultant predicted cases were assigned to each county accordingly and stored as a dictionary. Next, the dictionary was sorted according to its values of predicted cases, from highest to lowest, and the new sorted dictionary keys were assigned to a list. That list now contains counties’ names sorted from highest to lowest according to risk of pertussis outbreaks based on predicted number of cases. Lastly, the list is compared to the rates data CSV file on the 2014 year’s cases and an accuracy score was calculated.

The actual execution of creating the first model started with reading the student data and the pertussis rates CSV files, as both CSV files were encoded using ISO-8859-1, it presented a challenge, and the files were not read, but the appropriate parameter of the ISO-8859-1 encoding was specified in the read operation, and the files were read properly. Next, the student data CSV file was inspected. Upon inspection, the data cleaning process started. This consisted of five steps that cleaned the reshaped the data in order for the data to be properly fitted into the model.

The next step in the creation and execution of the first predictive model is the actual formation of the linear regression model. Here, the model fitted the created data frame containing the vaccinated and unvaccinated averages as pairs to represent the x-axis of the model (to represent the input) and the case’s average as the y-axis (to represent the output). Next, the model used cross validation with a 70/30 training to testing split ratio for the training process. Lastly, the model was used to predict on the data frame containing the 2014 vaccinated and unvaccinated data, where the output was formatted as a list representing number of predicted cases for each county. As the model used linear regression, the predicted values were continuous (contains floats) so all the values were rounded as needed. The list was then mapped to a dictionary with a key to value pair, where each key had the county’s name (in same order as original data frame) and each value had the predicted values. The dictionary was then sorted according to values from highest to lowest and the resulting ordered keys were stored as a list representing the counties sorted from highest to lowest risk of pertussis outbreaks.

The final step in the creation of the first model deals with the testing and validation process. This started with the reordering of the counties in the rates data CSV file for the 2014 year according to highest to lowest cases and storing the result of the ordered counties as a list. Then an accuracy score value was calculated on the two lists of the predicted values to actual values of the 2014 year. The resulting accuracy score was 12%, which was enough to predict the two counties with highest risk being Los Angeles and San Diego accurately, as well other counties of high risk in proper order. However, the overall accuracy of the first model was low as it failed to predict many of the other counties in California. This demanded the creation of a second, more improved model.

The creation of the second predictive model involved classification. As classification requires categories of data to classify against, the first step involved classifying pertussis cases values into three categories: high, medium, and low. High classification of pertussis cases involved number of cases from 140 and up, and high classification was mapped to a value of 3. Medium classification involved number of cases between 10 and 139 inclusive, and was mapped to a value of 2. Lastly, low classification involved number of cases between 0 and 9 inclusive, and was mapped to a value of 1. All the ranges used for the classification were chosen according to the ranges of cases of the year 2014. The data cleaning process for the second model is similar to the first model, but with some minor changes. Since the input for classification (x-axis) used in the model represents single values instead of pairs, the vaccinated averages were dropped and only the unvaccinated averages were kept. Next, a looping structure was performed on the updated student data frame and the cases averages for each county was updated and changed from numerical to categorical, where every numerical value was converted to the appropriate classification (1, 2, or 3). Next, a decision tree classifier was used to fit the data frame and the x-axis represented single value items of unvaccinated averages from the 2010 to 2013 years, while the y-axis represented cases with categorical classification instead of numerical. The model was fitted on the data with a 70 to 30 training to testing ratio split. The resulting model was then used to predict on the 2014 data of unvaccinated students, and the output was stored as a list which represents a series of 1s, 2s, and 3s for each county in the same order as original data frame. A dictionary was created with a key to value mapping to that list.

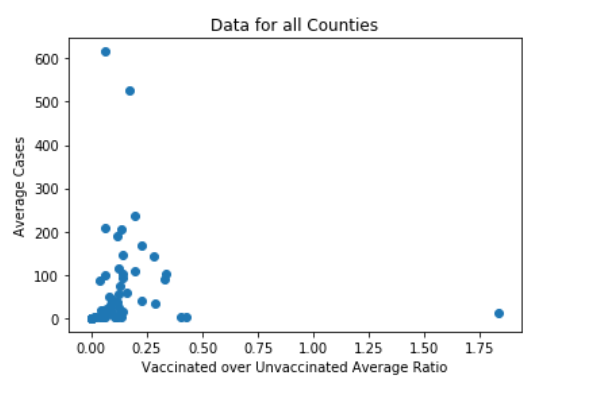
Lastly, for the testing and validation process of the second model, the rates data of actual cases for each county for 2014 year needed to be converted from numerical values to categorical representation for proper comparison. A similar process of changing the numerical values to categorical classification based on the previous rules of ranges was used. Then, another dictionary was created with a mapping of the county to categorical classification to represent actual values, and an accuracy score was calculated based on the predicted classifications to the actual classifications for every county. The resulting accuracy score was 67%, which is much more accurate and efficient than the first model.

**Conclusion**

In conclusion, exploratory data analysis helped in answering the questions provided in this project, with regards to the relationship between public/private vaccinations. Private schools have lower vaccination rates compared to public schools over the years. Furthermore, a negative correlation was shown between vaccination rates and personal belief exemption, as a result, personal belief exemption played a large role in lowering pertussis vaccinations. In addition, a linear relationship was shown between DTP, MMR, and polio vaccination over the years. Lastly, many of the counties displayed below and at a 90% vaccinated population which helped in spreading pertussis cases especially in the 2014 year.

**Future Work**

As to the predictive model portion of this project, increase in accuracy defines future work. While having access to more data without the presence of too many outliers can already increase the accuracy with the current predictive model, utilizing a different predictive model can lead to possibly improved efficiency. One possibility is utilizing curve fitting to better fit the shape of the data and create more accurate projections from which to predict. A quick exploration of the possibility of such a model involved the plotting of the ratios of vaccinated vs unvaccinated to represent the x-axis as single items, and the cases' averages to represent the y-axis.



Observing the plot (shown above), a mathematical function can be created to find the best fitted curve and represent the new predictive model. However, one possible issue with such model can be overfitting. To prevent such issue it demands the need for multiple mathematical functions and comparisons of those functions to choose the best one.

**References**

[1] Alamaw, Sefi Derib, et al. “Pertussis Outbreak Investigation of Mekdela District, South Wollo Zone, Amhara Region, North-West Ethiopia.” BMC Research Notes, vol. 10, no. 1, 2017, doi:10.1186/s13104-017-2735-1.

[2] Shute, Nancy. “Vaccine Refusals Fueled California's Whooping Cough Epidemic.” NPR, NPR, 30 Sept. 2013, www.npr.org/sections/health-shots/2013/09/25/226147147/vaccine-refusals-fueled-californias-whooping-cough-epidemic.

[3] Winter, Kathleen, et al. “Pertussis Epidemic - California, 2014.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 5 Dec. 2014, www.cdc.gov/mmwr/preview/mmwrhtml/mm6348a2.htm.

[4] Eggertson, L. “Lancet Retracts 12-Year-Old Article Linking Autism to MMR Vaccines.” Canadian Medical Association Journal, vol. 182, no. 4, 2010, doi:10.1503/cmaj.109-3179.

[5] Helft, Laura, and Emily Willingham. “What Is Herd Immunity?” *PBS*, Public Broadcasting Service, 5 Sept. 2014, [www.pbs.org/wgbh/nova/article/herd-immunity/](http://www.pbs.org/wgbh/nova/article/herd-immunity/).

[6] Awaidy, Salah T. Al. “Strategic Approaches Towards Pertussis Control in Oman.” *Oman Medical Journal*, vol. 33, no. 1, 2018, pp. 29–36., doi:10.5001/omj.2018.06.

[7] Korkmaz, Huseyin Anil, et al. “Comparison of Acellular Pertussis-Tetanus-Diphtheria Vaccines and Whole-Cell Pertussis-Tetanus-Diphtheria Vaccines in Infancy.” Paediatrics and International Child Health, vol. 34, no. 3, 2013, pp. 198–202., doi:10.1179/2046905513y.0000000110.

[8] Lehman, Deborah. “Pertussis Outbreak in California: Trends and Implications for Prevention.” NEJM Journal Watch: Summaries of and Commentary on Original Medical and Scientific Articles from Key Medical Journals, [www.jwatch.org/na36499/2014/12/16/pertussis-outbreak-california-trends-and-implications](http://www.jwatch.org/na36499/2014/12/16/pertussis-outbreak-california-trends-and-implications).

[9] Patterson, Jenna, et al. “Comparison of Adverse Events Following Immunisation with Acellular and Whole-Cell Pertussis Vaccines: A Systematic Review.” *Vaccine*, vol. 36, no. 40, 2018, pp. 6007–6016., doi:10.1016/j.vaccine.2018.08.022.

[10] Plans, P., et al. “Effectiveness of Acellular Pertussis Vaccination during Childhood (<7 Years of Age) for Preventing Pertussis in Household Contacts 1–9 Years Old in Catalonia and Navarra (Spain).” *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 35, no. 12, 2016, pp. 2059–2067., doi:10.1007/s10096-016-2766-x.