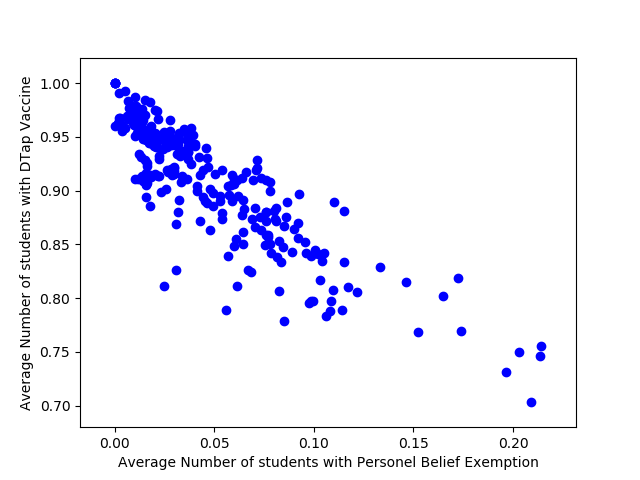
**Data Analysis**

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 and used Random Forrest Classifier 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.

**Exploratory Data Analysis**

First the team made pairplots of the data. Pairplots build histograms and scatter plots. The histograms show the distribution of a single variable and the scatter plots shows the relationship between two variables. From this plot, we are able to see what has a relationship and what doesn’t. The pairplot shows a linear relationship between the DTP(Diphtheria/Tetanus/Pertussis) vaccine and MMR(Measles/Mumps/Rubella) and Polio vaccine. These variables have a linear relationship because if a student gets the pertussis vaccine they most likely are going to get the other vaccines. We also saw there was a negative correlation between the number of students who got the pertussis vaccine and the number of students who claimed to have a personal belief exemption.



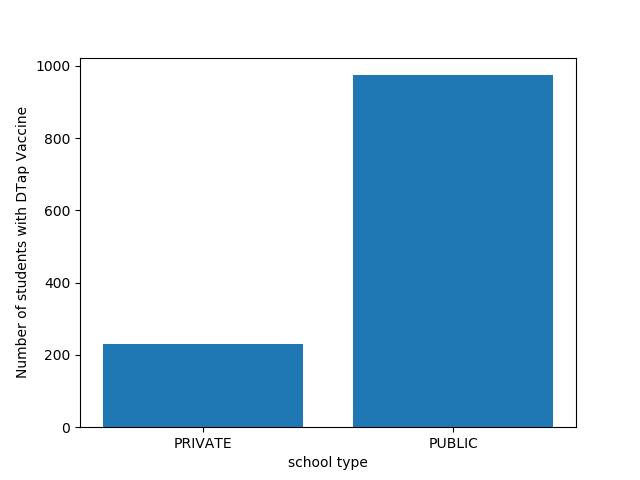
**Figure 1:** Scatter plot of number of students who got pertussis vaccine and number of students who claim personal belief exemptions between 2010-2014. (Need fix plot)

Some question I had about the data were

Do Public Schools or Private Schools have higher outbreak rates?

Are Vaccination rates increasing or decreasing?

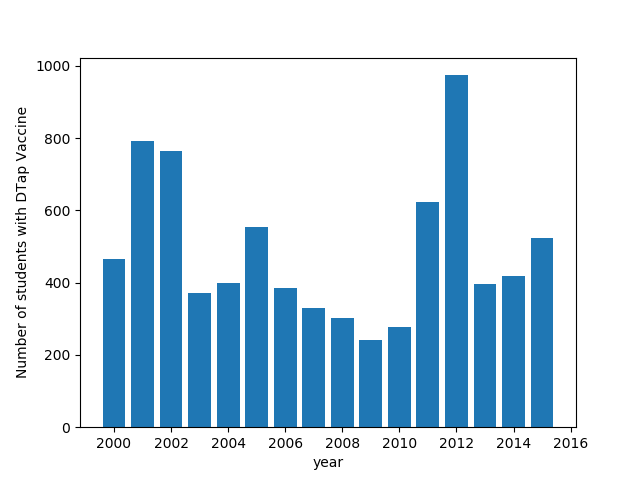
Which county has the highest outbreak rate?



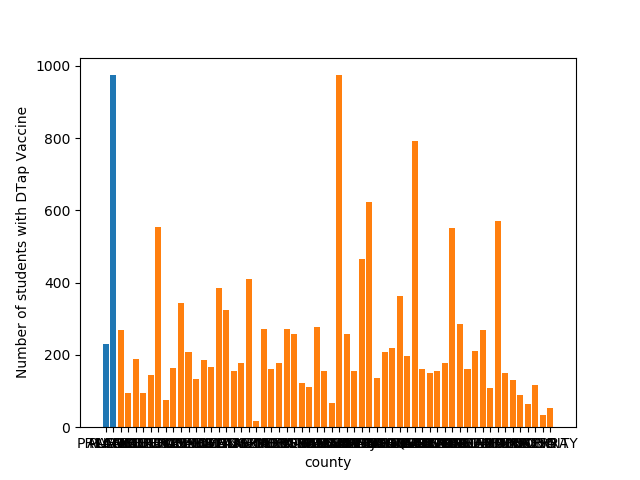
**Figure 2:** Bar chart shows public school children are getting the vaccine more than private school children.

Children in private schools are not getting vaccinated as much as public school children.

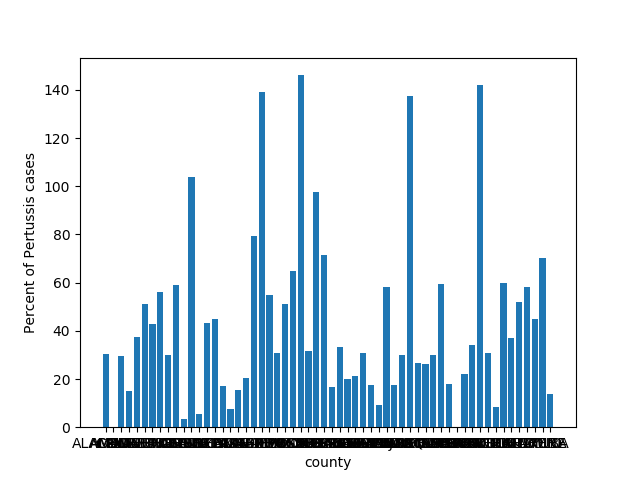
We decided to plot the number of students who got the pertussis vaccine for the years 2000-2014 to see if there was any trend. The first low point in the data we have is in 2000. This is due to a paper published in 1998 in the Lancet. The paper implied a link between vaccinations and autism. In the original paper, Wakefield and 12 coauthors claimed to have investigated “a consecutive series” of 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 “possible environmental triggers” (i.e. the vaccine) were associated with the onset of both the gastrointestinal disease and developmental regression. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831678/> ) The paper was first investigated by the Lancet in 2004. Twelve years later the paper was retracted but many autism advocacy groups and parents continue to defend the papers findings. We found that 2009-2010 had some of the lowest vaccination rates. This is due to parents claiming personal belief exceptions. 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. (<https://www.npr.org/sections/health-shots/2013/09/25/226147147/vaccine-refusals-fueled-californias-whooping-cough-epidemic> ) Figure 5 we saw that the vaccination rate varied a lot by county and that some counties had a lot more cases. (Clara talks more about this)



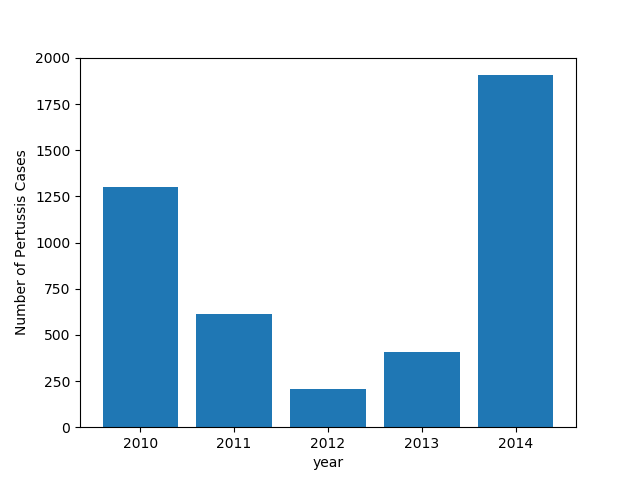
**Figure 3:** Shows the number of students with pertussis vaccination between 2000-2015



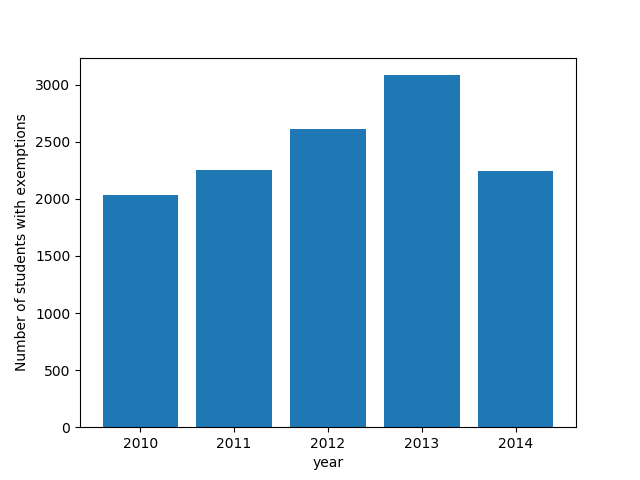
**Figure 4:** Number of students who got DTaP vaccine in each county. (Need fix plot)



**Figure 5:** (Need fix plot).



**Figure 6:** Total Pertussis cases for the years 2010- 2014



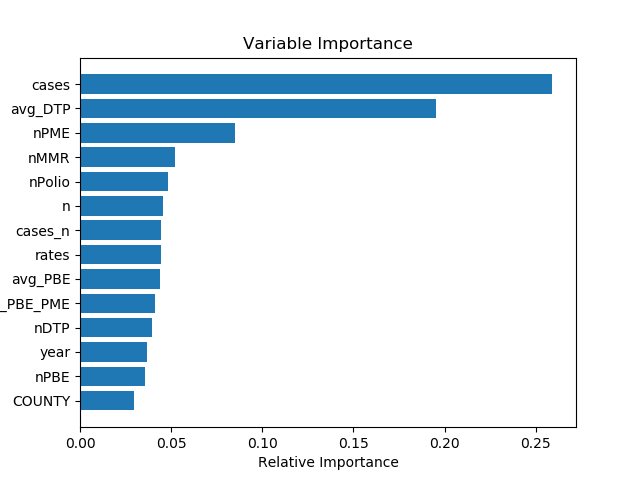
**Figure 7:** Total number of personal belief exemptions

What happened in 2014?

In September 2010, a bill was signed requiring students entering 7th grade to show proof of immunization with a pertussis booster. The number of pertussis cases in 2014 was larger than in 2010. In 2014, there was an increase in cases of 12-17 year olds. (source)

**Random Forrest Classifier**

Random Forrest classifier is a method for classification and regression. We used the feature\_importances\_ attribute of this classifier in addition the data analysis to figure out which variables affect the number of pertussis cases the most.  This information was used to drop out features that look like noise. To build the model I set the target variable to be pertussis rates for 2010-2014 and plotted the results in a bar chart. From the model and the other data analysis the top three factors are number of pertussis cases, number of students with DTap vaccine and number of exemptions.



**Figure 8:**