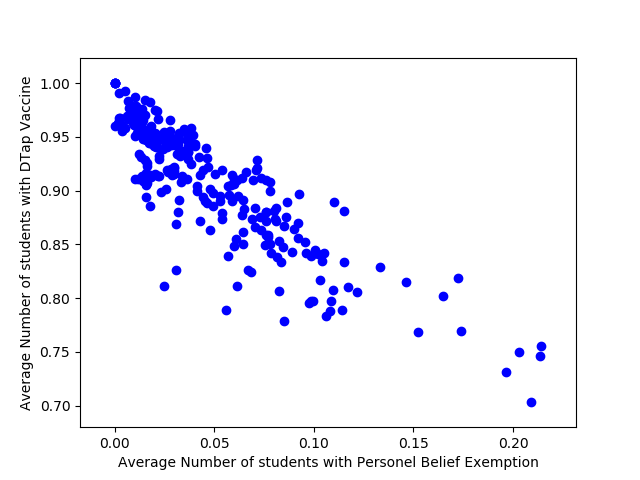
**Data Analysis**

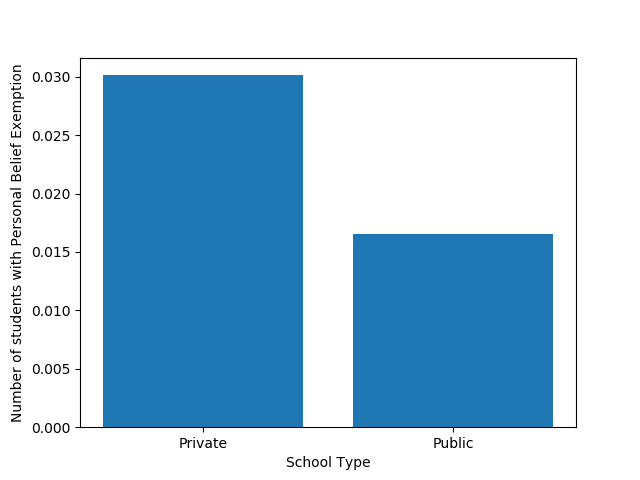
For the data analysis portion of the project we did exploratory data analysis and kMeans clustering to find trends in the data. We made plots and used Random Forrest Classifier to identify key features that affect the number of pertussis cases. Using kMeans clustering we found there was a possible relationship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

**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 most likely have a linear relationship because if a student gets the pertussis vaccine they are going 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 have a personal belief exemption.

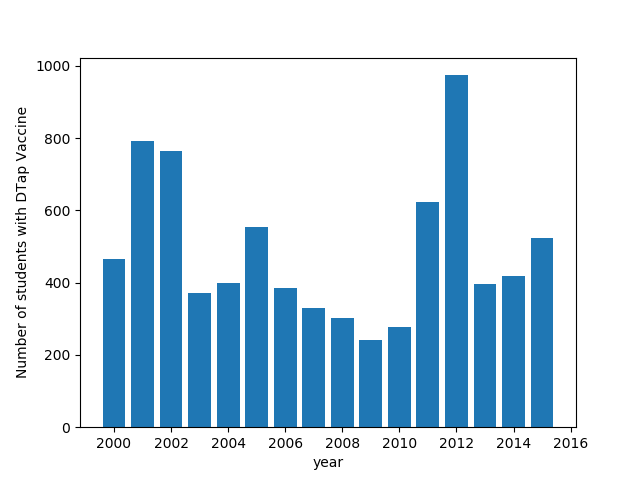


**Figure 1:** Scatter plot of number of students who got pertussis vaccine and number of students who claim personal belief exemptions between 2010-2014.



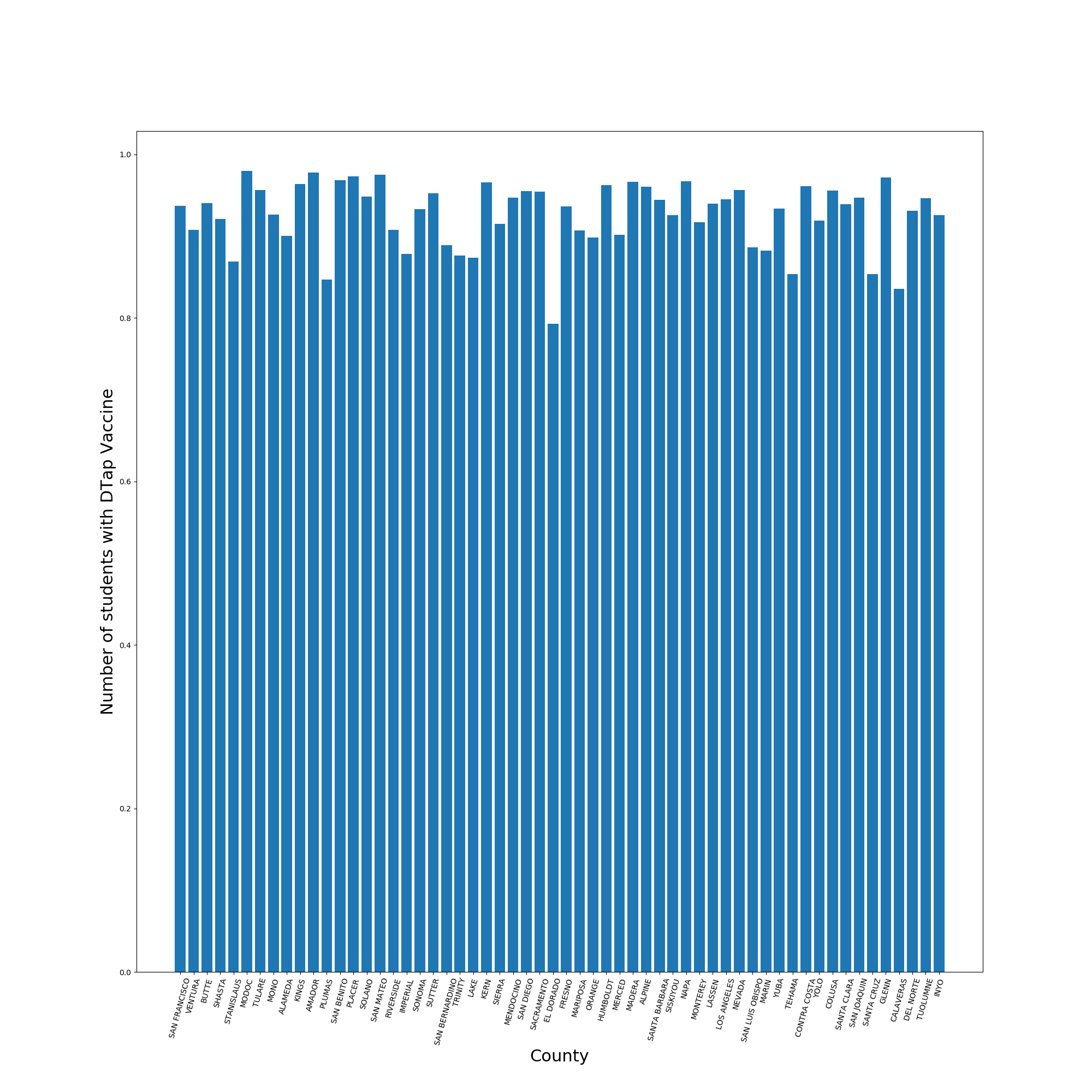
**Figure 2:** Bar chart shows private school children claim a personal belief exemption more than public school children.

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. (<https://www.npr.org/sections/health-shots/2013/09/25/226147147/vaccine-refusals-fueled-californias-whooping-cough-epidemic> )



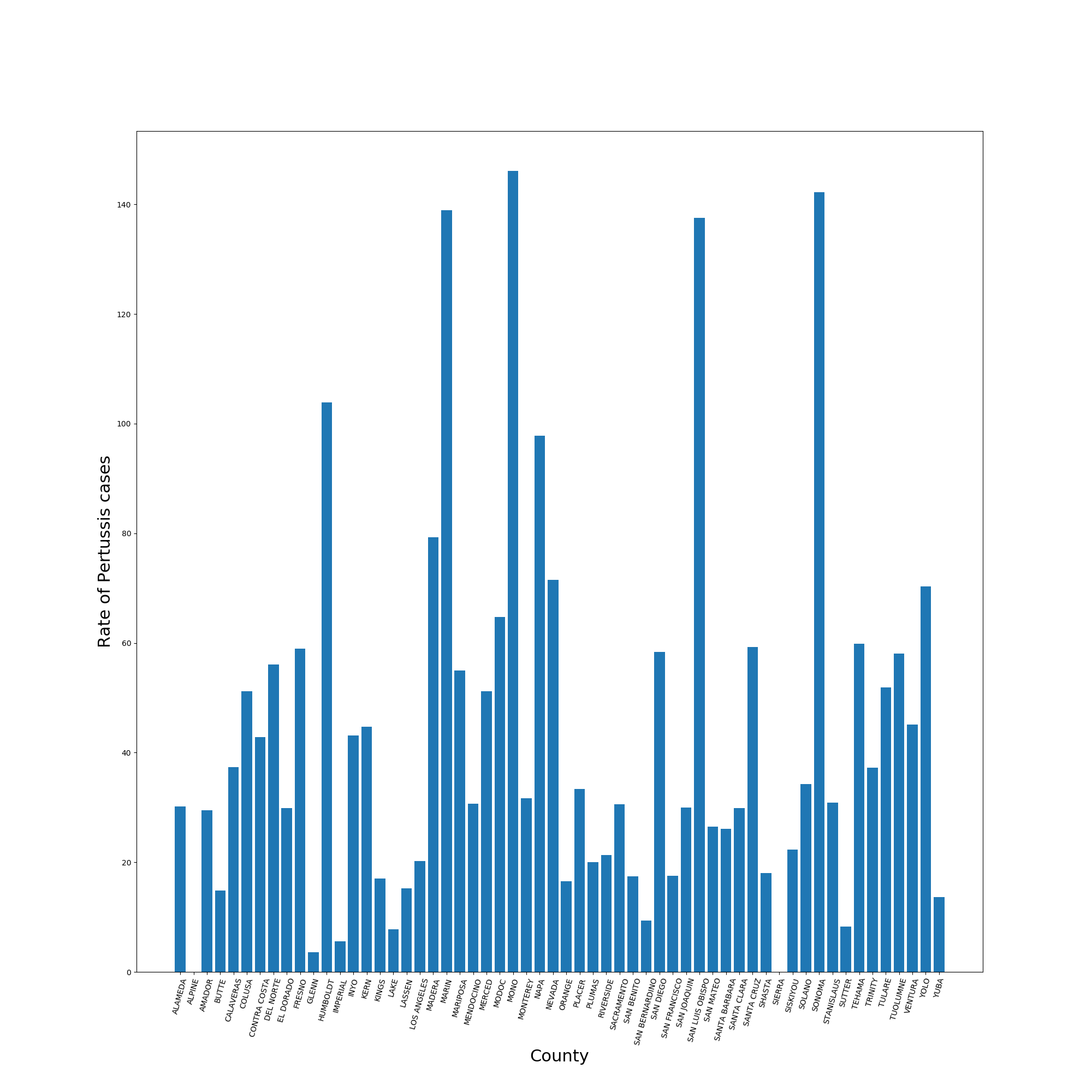
**Figure 3:** Shows the number of students with pertussis vaccination between 2000-2015

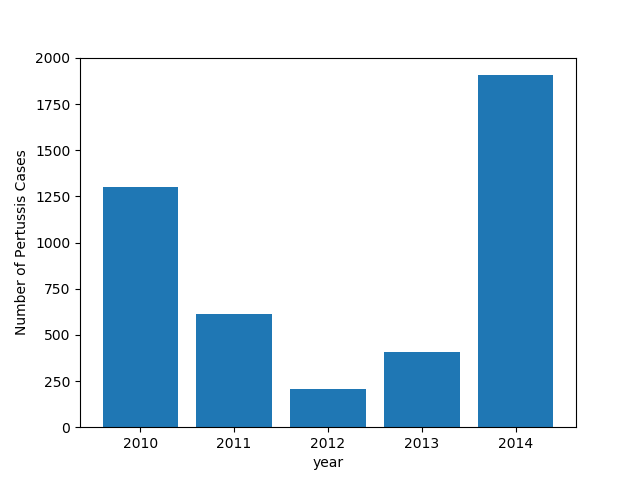
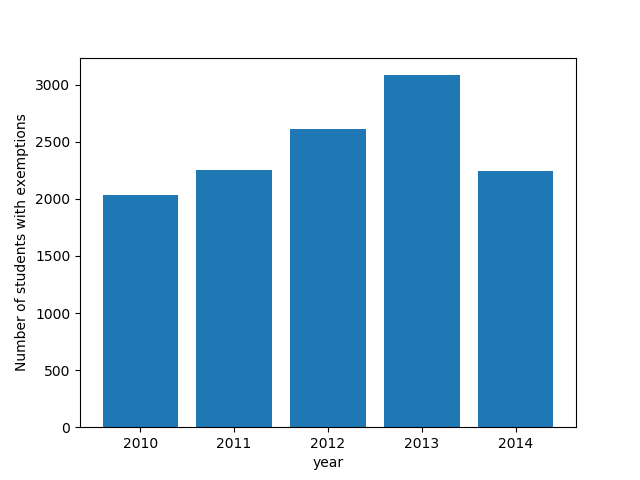
We decided to plot the number of students who got the pertussis vaccine for the years 2000-2014 to see if there was any trend. The first low point in the data we have is in 2000. This is due to a paper published in 1998 in the Lancet. The paper implied a link between vaccinations and autism. In the original paper, Wakefield and 12 coauthors claimed to have investigated 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. (<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 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 95% then herd immunity is lost. (source)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 that area. Figure 5 we saw that the rate of pertussis cases varied a lot by county and that some counties had a lot more cases. The rate is the incident rate for every 100,000 people.





**Figure 5:** Rate of pertussis cases per 100,000 persons in each county

**Figure 6:** Total Pertussis cases for the years

2010- 2014

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

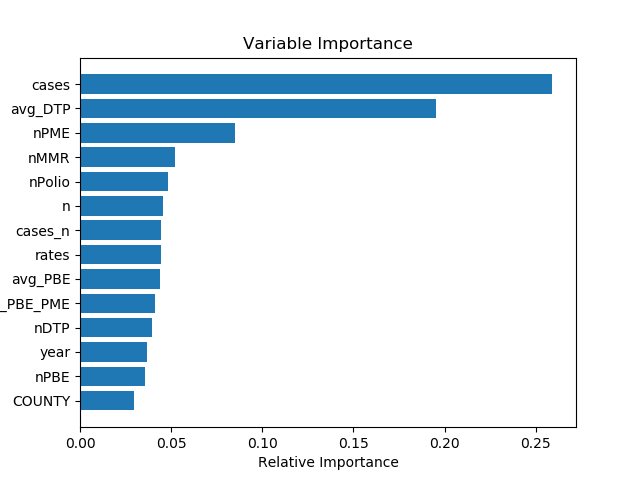
What happened in 2014?

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. (<https://www.cdc.gov/mmWr/preview/mmwrhtml/mm6348a2.htm> ) 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. ( <https://www.cdc.gov/mmWr/preview/mmwrhtml/mm6348a2.htm> )

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 infants. They were the first to receive acellular pertussis vaccines, used exclusively since 1997 and considered to have less-durable immunity. (<https://www.jwatch.org/na36499/2014/12/16/pertussis-outbreak-california-trends-and-implications> ) 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 and 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. In September 2010, a bill was signed requiring students entering 7th grade to show proof of immunization with a pertussis booster.

**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:** Variable Importance