Capstone Project: How to Manage Diseases?

By:

Wilson Aleman

Colin Beran

Scott Cummings

Rishi Paul

Allen Tung

ISDS 577 - Capstone Project

Spring 2020

California State University, Fullerton

**Table of Contents**

[**Overview**](#_u0xenss65h1w) **5**

[**Dataset**](#_yqyybbpwc8ic) **6**

[**Data Transformation**](#_3d1ppqx7oen) **7**

[**Question 1: Does a location’s economic situation affect the rate of immunization among school children?**](#_yqrhlj9lxtcb) **7**

[**Question 2: Does DTP vaccine exemption affect case rates in California?**](#_d516b7ylm7ly) **8**

[**Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?**](#_y1m7uprten5z) **9**

[**Results**](#_d77jfhbe9c0y) **10**

[**Question 1: Does a location’s economic situation affect the rate of immunization among school children?**](#_t43uct5iwms0) **10**

[Exhibit 1-1: Bar Chart Comparison of Immunization Rate and Poverty (Under Age 18)](#_v19pgau4g5o8) 11

[Exhibit 1-2: Bar Chart Comparison of Immunization Rate and Poverty](#_tio3my6s50eh)  12

[Exhibit 1-3: Bar Chart Comparison of Immunization Rate and Population](#_3rwke85o396g) 13

[Exhibit 1-4: Bar Chart Comparison of Immunization Rate and % of Adults with a Bachelor’s Degree or Higher](#_ah7yfbgmgozs) 14

[Exhibit 1-5: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Poverty (Under Age 18)](#_6ry1xi4olplt) 16

[Exhibit 1-6: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Poverty (All Ages)](#_1bx9rebzqy) 17

[Exhibit 1-7: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Population](#_mto2c1d40nzx) 19

[Exhibit 1-8: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and % with Bachelor’s Degree or Higher](#_mkd74nltjnsv) 21

[Exhibit 1-9: Model Summary and Coefficients of Multiple Regression Model](#_czx7vwxuh04p) 22

[**Question 2: Does DTP vaccine exemption affect case rates in California?**](#_s6h7m2zv4by) **23**

[Exhibit 2-1: Amount of school data per year.](#_jmax0u6r1p03) 24

[Exhibit 2-2: Number of exemptions in the state of California from 2000 to 2015.](#_b3df4d20tnfb) 25

[Exhibit 2-3: Number of exemptions by county size in 2006, 2013, and 2014.](#_77o1kz992ycr) 26

[Exhibit 2-4: Box plots of normalized number of exemptions during 2006, 2013, and 2014.](#_yk8sny6xdhq) 27

[Exhibit 2-5: Counties with a stable average percentage of exemptions from 2006 to 2015.](#_egdaqnts0z4d) 28

[Exhibit 2-6: Counties with a non-stable average percentage of exemptions from 2006 to 2015.](#_bxoyq4fylkus) 28

[Exhibit 2-7: Graph of the top seven CA counties with the highest average percentage exemptions in 2014. Case rates in red are higher than the CA average during that same year.](#_5cxu1fcietkr) 30

[Exhibit 2-8 - Graph of CA counties with zero average percentage exemptions in 2014. Case rates in red are higher than the CA average during that same year.](#_xcrntmi8bmxr) 30

[**Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?**](#_vurvozz9z3uq) **31**

[Exhibit 3-1: SPSS results from the initial model using all potentially relevant variables.](#_yn9qr81ssdom) 32

[Exhibit 3-2: SPSS results from the improved model using quadratic regression.](#_iq4xgdwvue2s) 33

[Exhibit 3-3: Regression Tree for determining the number of cases.](#_8z04a3kvyyqw) 35

[Exhibit 3-4: Minimum Error Regression Tree for determining the number of cases.](#_aadkmhhmlq4h) 35

[Exhibit 3-5: Map of 2010 Pertussis cases among kindergarten students by California county.](#_22giv8y98gj8) 36

[**Conclusion**](#_f3sl98dh1890) **37**

[**Question 1: Does a location’s economic situation affect the rate of immunization among school children?**](#_vc7linyho0a3) **37**

[**Question 2: Does DTP vaccine exemption affect case rates in California?**](#_uexj178oof1s) **38**

[**Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?**](#_ehqaf8vyxqbl) **39**

[**Appendices**](#_f3sl98dh1890) **40**

[**Appendix A: SQL Code**](#_iqbv6lw5jap) **40**

[SQL query to combine and create CountyRates table, listing each element as a sum by year and county.](#_f8nviaibhaw2) 40

[**Appendix B: R Code**](#_iqbv6lw5jap) **41**

[R script used to create a table to average percentage exemptions for all California counties](#_n239mxj3qkhi) 42

[Script used for creating regression trees.](#_7s90ksgj5udj) 43

# Overview

Only a few months ago, at the start of 2020, the American public was uninterested and unprepared for an epidemic like the vicious Coronavirus Disease (COVID-19). This epidemic has brought worldwide attention to the dangers of a fast-spreading disease. Daily, the public is being told by the news what the Center for Disease Control (CDC) has recommended doing: washing hands, avoiding close contact with others, and cleaning and disinfecting surfaces.

COVID-19 has caused many people to lose jobs, houses, and the freedom to be with others. In response to the terrible consequences of this virus, it is vital to analyze more closely how diseases spread and what factors affect them. Our group decided to focus our efforts on analyzing a dataset from Kaggle called California Immunization Rates, and another dataset from the USDA Economic Research Service (ERS). There are no significant legal or privacy concerns to use these two datasets. The data on Kaggle is an amalgam of publicly available data compiled or converted from the

California Department of Public Health. The USDA ERS data is also publicly available and anonymized.

Three questions are answered to decision-makers about how to manage highly contagious diseases: Does a location's economic situation affect the rate of immunization among school children? Does Diphtheria, Tetanus, Pertussis (DTP) vaccine exemptions affect case rates in California? Which California counties and schools have the highest risk of facing a Pertussis outbreak? The answer to the first question may provide valuable context to government agencies and health-focused NGOs on which population groups that are most at risk for vaccine-preventable diseases. Ideally, this may lead to more targeted outreach and intervention that can prevent or mitigate future outbreaks. The second question addresses one of the World Health Organization's top global health threats, which is anti-vaccine movements. Decision-makers need to understand the effect of immunization exemptions in schools. Answering the third question may identify which counties and schools are at the highest risk of diseases such as Pertussis to take precautionary measures (i.e., hand sanitizing stations, and social distancing).

# Dataset

The primary datasets were California Kindergarten Immunization Rates and Pertussis Cases. These records are publicly accessible through links from Kaggle, an online community of data scientists and machine learning practitioners, containing student immunization data from over 7000 public and private schools in California between the years 2000 to 2015. There were three types of vaccinations tracked in the data - Measles, mumps, and rubella; polio; and diphtheria, Pertussis, and tetanus. Also available was the number of students that fell under personal belief exemptions and permanent medical exemptions for each school. A separate file contained data on pertussis cases between the years 2010 to 2014 by California county.

Additional datasets are from publicly available sources from the USDA Economic Research Service (ERS) and the US Census Bureau. County-level data regarding percentage in poverty, population, and education levels were obtained from these sources. Despite an additional effort, no consistent datasets appeared to be available for similar data presented at a more granular level, such as city/township or zip code.

# Data Transformation

## Question 1: Does a location’s economic situation affect the rate of immunization among school children?

This question required the use of California Kindergarten Immunization Rates and county-level data from the USDA ERS and US Census Bureau. The period selected to answer this question was 2010, as it appeared to have the most extensive data sets among all the sources and still relatively close to the present. The only data used that did not match this period was county-level education levels, which did not have complete data for 2010. Instead, the 2014-18 American Community Survey five-year average county-level estimates were used as it was the closest approximation.

The California Kindergarten Immunization Rates provided information by individual schools. The dataset required processing to use in conjunction with the county-level data. Microsoft Excel was used to clear irrelevant columns and to filter out and remove all rows that were not 2010. This partially transformed data was then imported into RStudio, where a script was run to aggregate the immunization data (number of kindergarteners immunized for measles, mumps, and rubella; polio; and diphtheria, Pertussis, and tetanus) into a single county percentage measure - see Appendix B.

## Question 2: Does DTP vaccine exemption affect case rates in California?

To obtain the number of exemptions per school, the California Kindergarten Immunization Rates dataset was used. The first transformation to this dataset was to add a county size attribute in order to analyze if there are any differences between Urban and Rural counties. The Rural County Representatives of California (RCRC) is an organization that champions policies on behalf of California's rural counties. Their website has a list of rural counties that are members, which helped in identifying rural counties. The other transformation was to create another attribute for normalized exemptions or percentage of exemptions. This attribute was calculated using the number of exemptions and dividing it by number of students at each school. Microsoft Excel was used for the first two transformations. However, to analyze the percentage of exemptions at the county level, we needed to group data by county then take an average. R-studio was used to create a table of the average percentage of exemptions by county from 2006 to 2010 - see Appendix B. This table was heavily used to see the effects of exemptions in the number of case rates for each county.

## Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?

This question required the use of the California Kindergarten Immunization Rates dataset as well as the Pertussis Rates dataset. The datasets required considerable cleaning and processing before analysis could begin, involving fixing clerical errors, miscellaneous data formatting differences, and data restructuring. The datasets were combined using SQL to have both the number of pertussis cases in each county and the information regarding kindergarten student immunization readily available. The SQL query used can be found in Appendix A. This new dataset was then imported into IBM’s SPSS software to perform regression analysis. After creating an initial regression model, all relevant factors (n, nDTP, nPME) from the immunization rates dataset required squaring to allow for quadratic regression modeling. After creating an acceptable quadratic regression model, both the data and the model were imported into R-Studio, where a script ran to create a regression tree. When satisfied with the quality of the regression tree, pruning generated a minimum error tree. The script used to create the trees is in Appendix B.

# Results

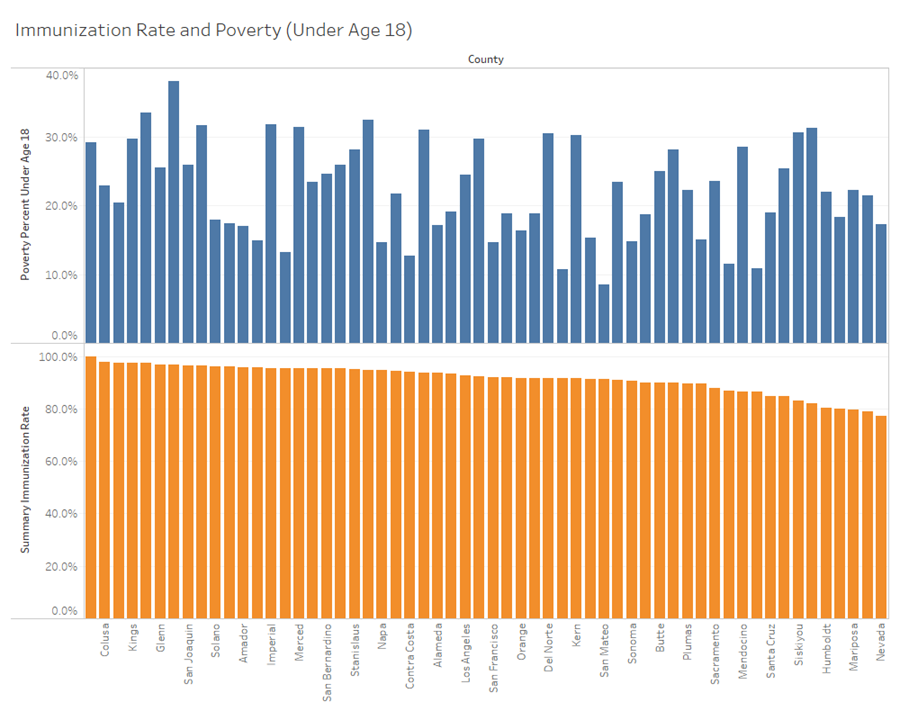
## Question 1: Does a location’s economic situation affect the rate of immunization among school children?

There is a general impression that more impoverished communities would tend to have lower immunization rates among their school children when compared to their better-funded counterparts. The testing of this rough hypothesis required performing simple linear regressions. The dependent variable was the aggregated county student immunization rate for each county of California. The independent variables were various economic and socio-economic measures.

Before running regressions, Side-by-side bar chart visualizations were created using Tableau to see if there were any easily discernible relationships between the variables. The immunization rate was sorted from highest to lowest from left to right and colored in orange. The corresponding variable is colored blue.

Visualization A

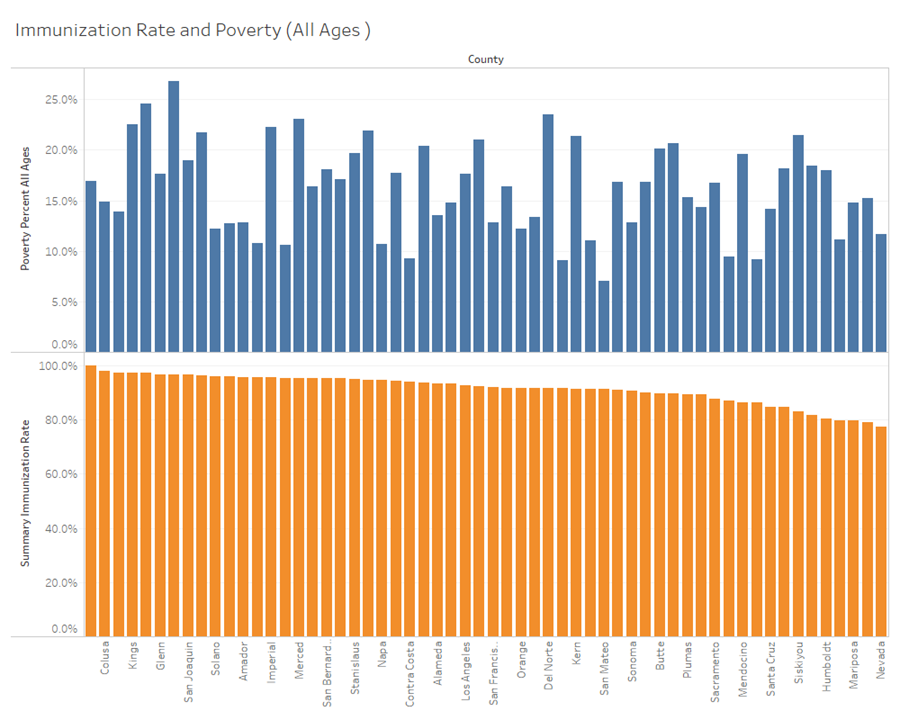
Immunization Rate and Percentage of Persons in Poverty Under Age 18



### Exhibit 1-1: Bar Chart Comparison of Immunization Rate and Poverty (Under Age 18)

Visualization B

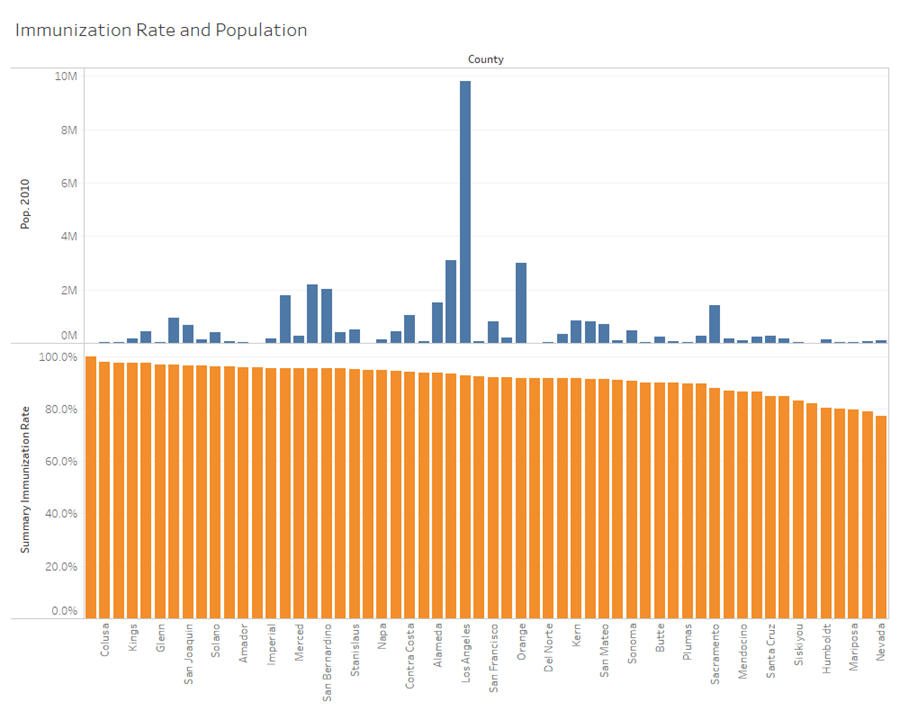
Immunization Rate and Percentage of Persons in Poverty, All Ages



### Exhibit 1-2: Bar Chart Comparison of Immunization Rate and Poverty (All Ages)

Visualization C

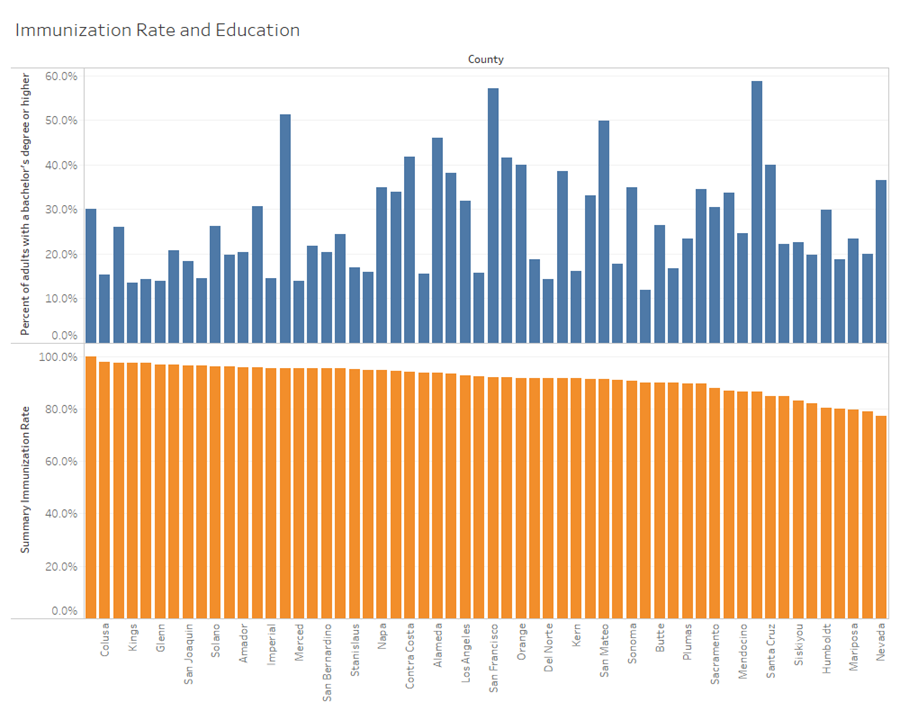
Immunization Rate and Population



### Exhibit 1-3: Bar Chart Comparison of Immunization Rate and Population

Visualization D

Immunization Rate and Percent of Adults with a Bachelor's Degree or Higher



### Exhibit 1-4: Bar Chart Comparison of Immunization Rate and % of Adults with a Bachelor’s Degree or Higher

Surprisingly, no apparent relationship arose using this visual comparison. However, this did not mean that there was no relationship and moved forward with running simple linear regression models. The variables and associated model summary and coefficients are presented below.

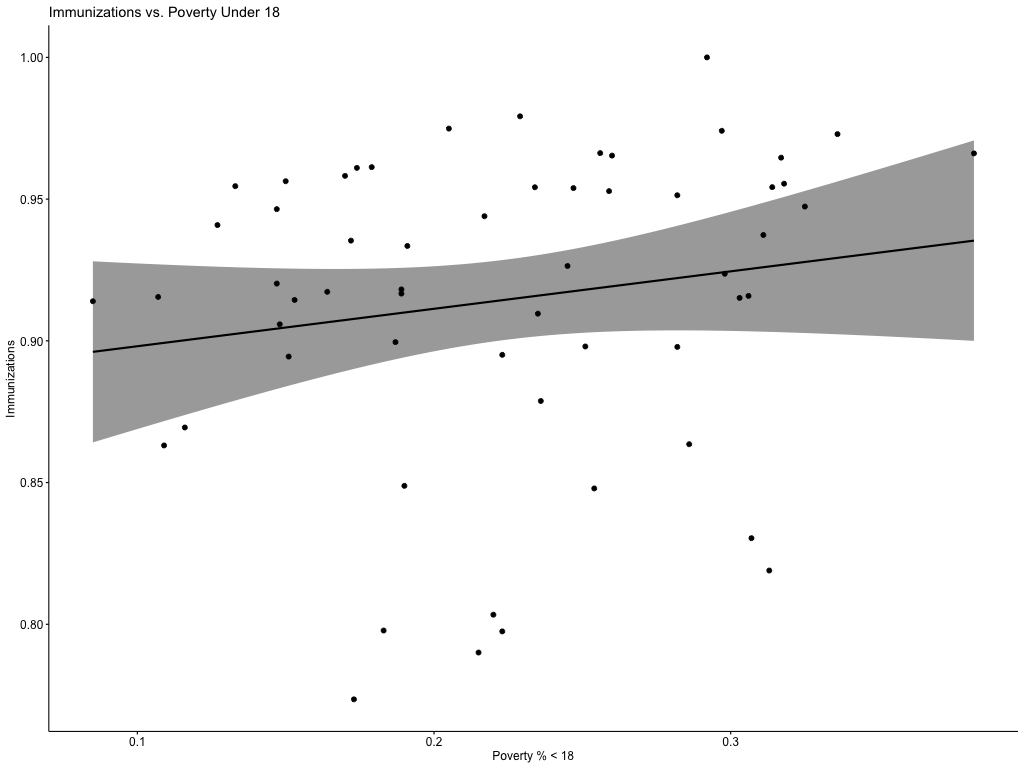
Model A

Dependent Variable: Immunization Rate

Independent Variable: Poverty Percent Under Age 18

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **R-Squared** | **Adjusted R-Sq** | **Std. Error** |
| 0.170 | 0.029 | 0.011 | 0.054 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficients** | **Coef Std. Error** | **t** | **Sig.** |
| **Constant** | 0.885 | 0.024 | 36.753 | 0.000 |
| **Poverty <18** | 0.132 | 0.103 | 1.288 | 0.203 |



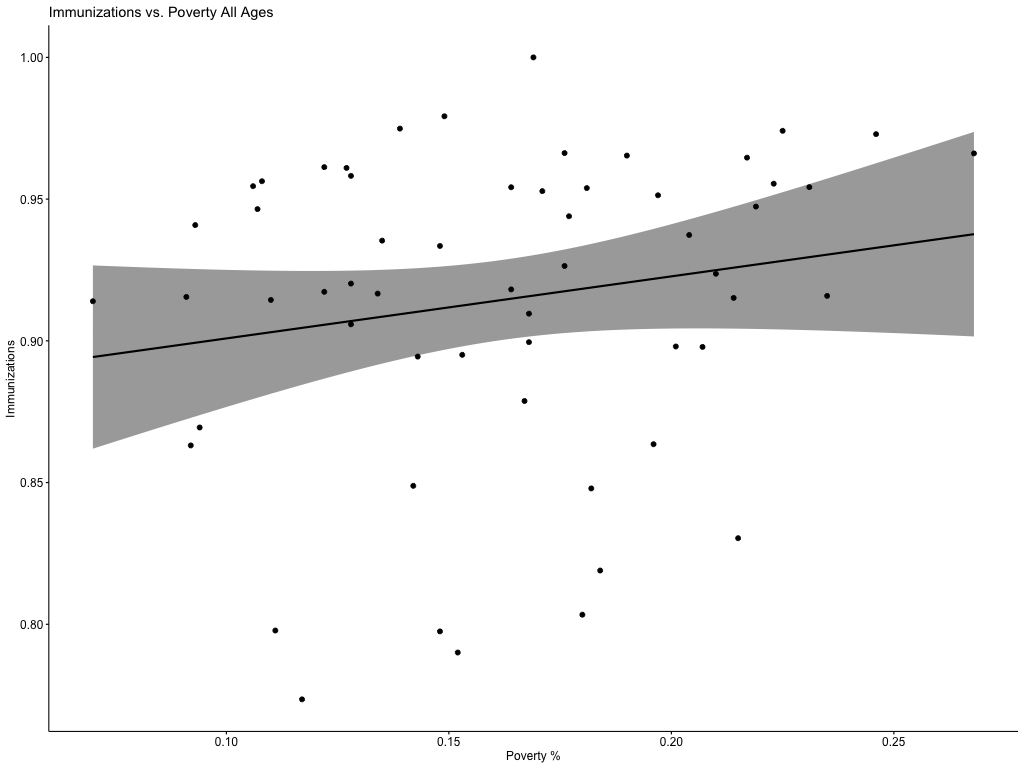
### Exhibit 1-5: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Poverty (Under Age 18)

Model B

Dependent Variable: Immunization Rate

Independent Variable: Poverty Percent All Ages

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **R-Squared** | **Adjusted R-Sq.** | **Std. Error** |
| 0.183 | 0.034 | 0.016 | 0.054 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficients** | **Coef Std. Error** | **t** | **Sig.** |
| **Constant** | 0.879 | 0.026 | 33.204 | 0.000 |
| **Poverty All** | 0.219 | 0.157 | 1.394 | 0.169 |

### Exhibit 1-6: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Poverty (All Ages)

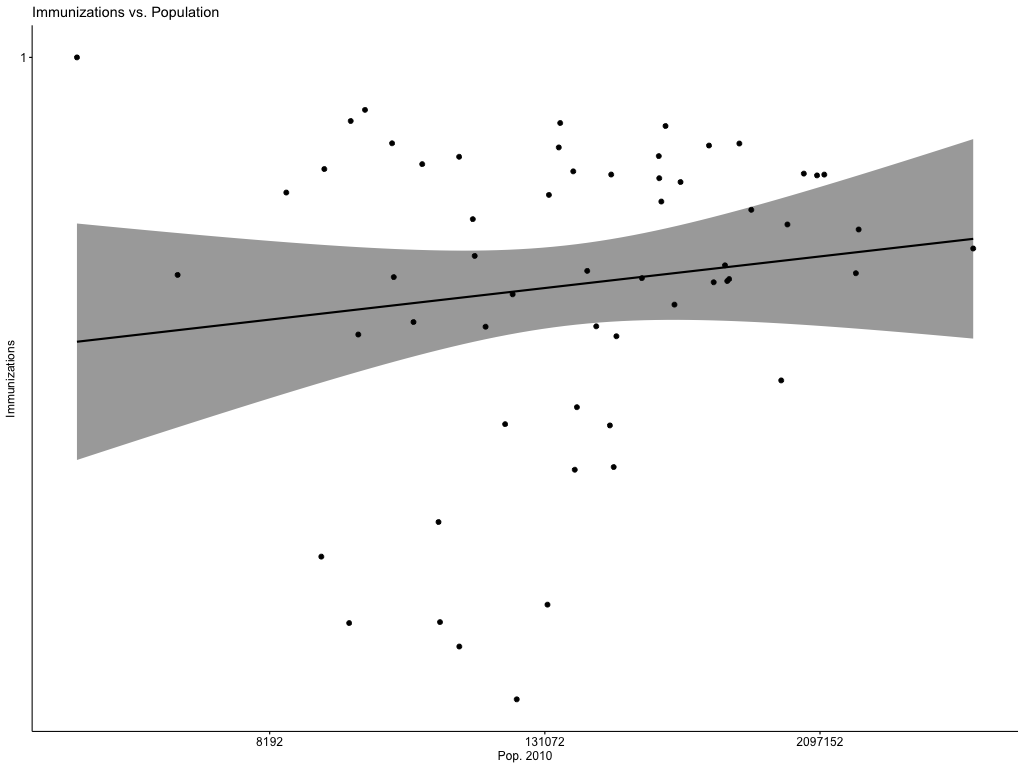
Model C

Dependent Variable: Immunization Rate

Independent Variable: Population

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **R-Squared** | **Adjusted R-Sq.** | **Std. Error** |
| 0.124 | 0.015 | -0.002 | 0.054 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficients** | **Coef Std. Error** | **t** | **Sig.** |
| **Constant** | 0.911 | 0.008 | 116.571 | 0.000 |
| **Population** | 0.000 | 0.000 | 0.936 | 0.353 |



### Exhibit 1-7: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and Population

\*County Population is displayed on a logarithmic scale due to the large population size of certain areas.

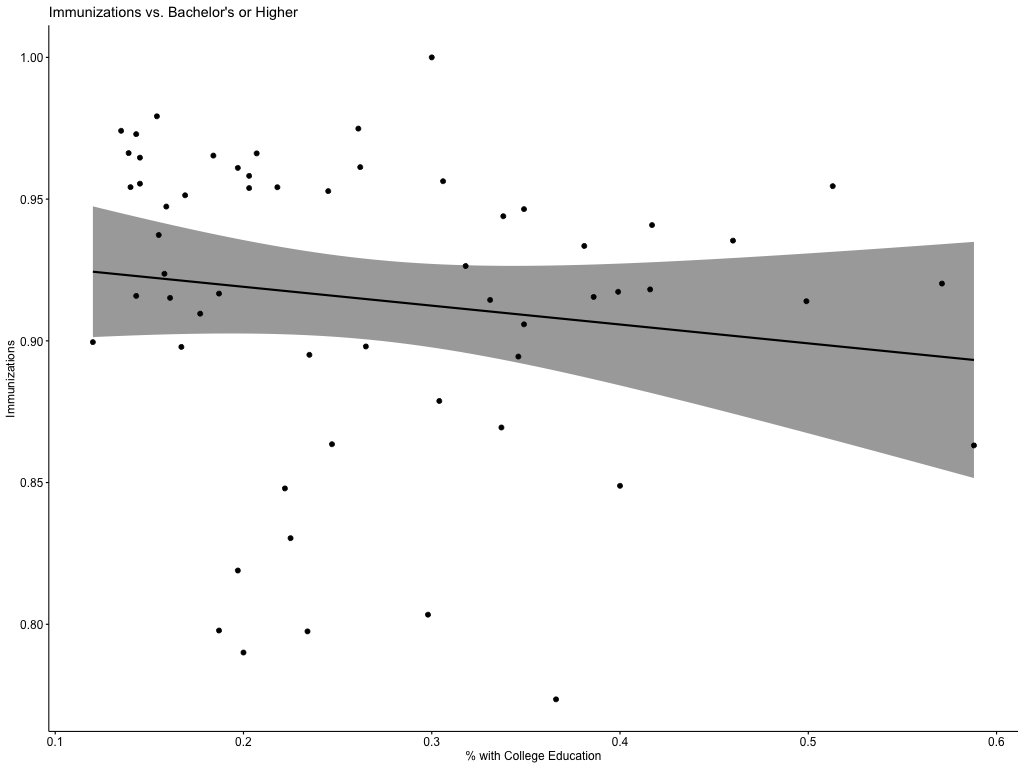
Model D

Dependent Variable: Immunization Rate

Independent Variable: Percent of adults with a bachelor's degree or higher

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **R-Squared** | **Adjusted R-Sq.** | **Std. Error** |
| 0.144 | 0.021 | 0.003 | 0.054 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficients** | **Coef Std. Error** | **t** | **Sig.** |
| **Constant** | 0.932 | 0.018 | 52.159 | 0.000 |
| **% with BA/BS or higher** | -0.067 | 0.061 | -1.088 | 0.281 |



### Exhibit 1-8: Model Summary, Coefficients, and Scatter Plot & Regression Line of Immunization Rate and % with Bachelor’s Degree or Higher

The results seemed to reflect the initial visual comparison - that no clear relationship could be linked to a single one of the predictor variables. The R-squared values in these simple regression models were all quite low - less than 0.05. Additionally, no predictor variable had a p-value below the typically accepted level of significance of 0.05. However, a multiple regression model would be created to see if a more substantial relationship would be evident when these variables were used in combination. Poverty under age 18 was left out to prevent excessive multicollinearity, and poverty at all ages was included since it had a comparatively higher R-square value.

Model E

Dependent Variable: Immunization Rate

Independent Variables: Poverty Percent All Ages, Population, and Percent of adults with a bachelor's degree or higher

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **R-Squared** | **Adjusted R-Sq.** | **Std. Error** |
| 0.235 | 0.055 | 0.003 | 0.054 |

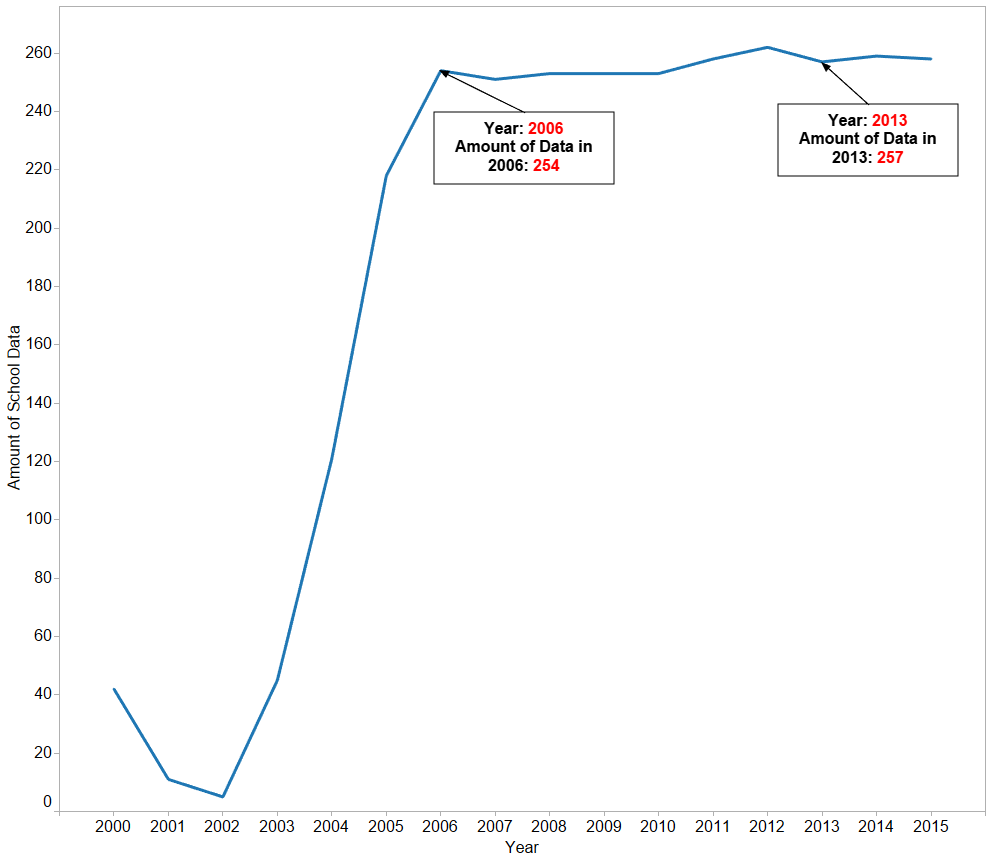
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficients** | **Coef Std. Error** | **t** | **Sig.** |
| **Constant** | 0.898 | 0.055 | 16.451 | 0.000 |
| **Poverty All** | 0.151 | 0.220 | 0.683 | 0.497 |
| **Population** | -0.043 | 0.088 | -0.490 | 0.626 |
| **% with BA/BS or higher** | 0.000 | 0.000 | 0.279 | 0.279 |

### Exhibit 1-9: Model Summary and Coefficients of Multiple Regression Model

Naturally, the models with additional variables will have a higher R-squared value. Therefore, the adjusted R-squared was used to compare with the simple regression models. It appears that there is little difference between this multiple regression model than with any of the other previous models. Similarly, the predictor variables have high p-values, which indicates that they are not significant.

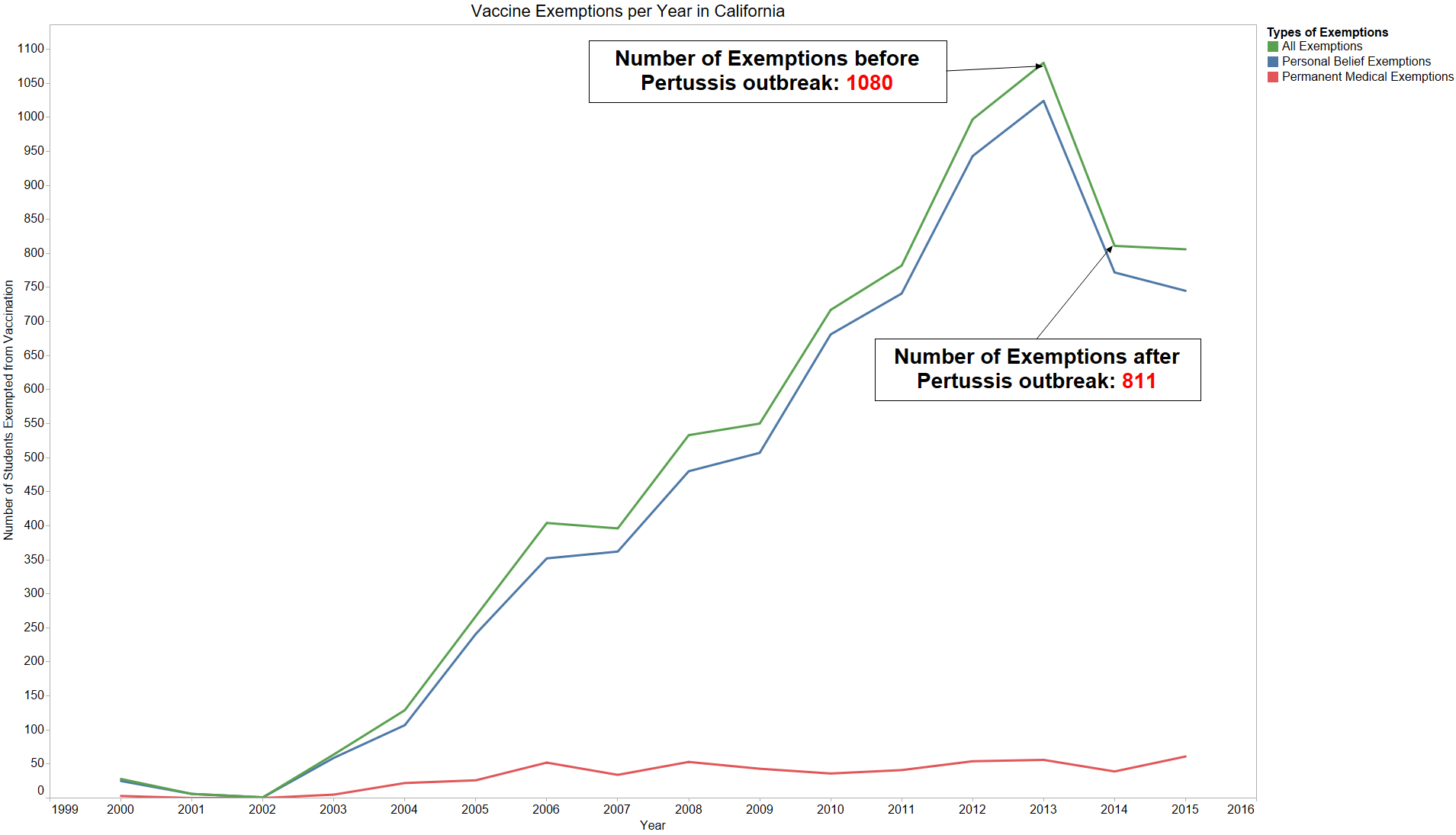
## Question 2: Does DTP vaccine exemption affect case rates in California?

Pertussis is a highly contagious bacterial disease, also known as whooping cough. Infants are at the highest risk for life-threatening cases of Pertussis. The latest outbreak of Pertussis in California occurred in 2014, causing several hospitalizations. Typically, these Pertussis outbreaks occur every three to five years. The state of California has taken measures such as Bill SB-277, which requires children to have diphtheria, pertussis, and tetanus (DTP) vaccination records. However, families still have the option to request an exemption not to vaccinate their child. An analysis was conducted on the dataset StudentData provided by Kaggle to understand the impact of these exemptions. The records in the dataset show the school's name, type of school (public or private), county, number of students, number of vaccinations, and number of exemptions. Exhibit 2-1 was the first graph of many created using Tableau to check which years were to be used for comparison purposes. This graph dictated that it would be appropriate to use data from the years 2006 to 2015 because those years have similar amounts of records.



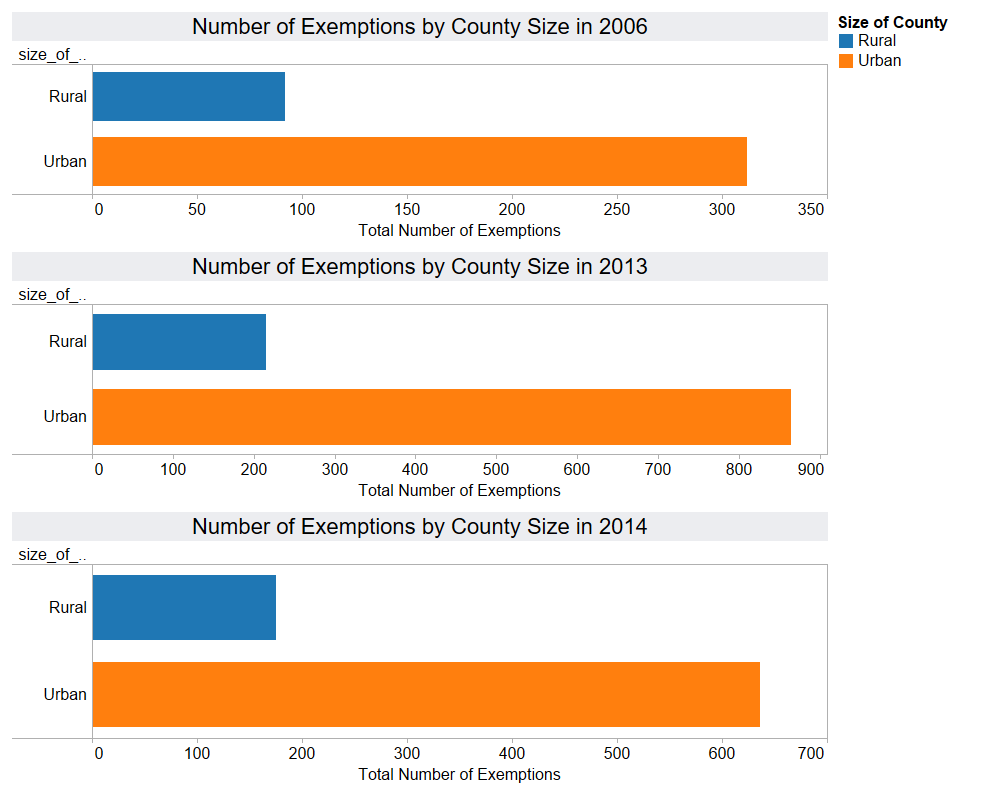
### Exhibit 2-1: Amount of school data per year.

Now, that year 2006 to 2015 were selected to compare other data. Another graph was developed to visualize the number of exemptions, Exhibit 2-2. From this graph, it is apparent that exemptions have increased in the state of California from 2000 to 2013. Personal beliefs and permanent medical exemptions were the only two types of exemptions recorded. The exemptions count is certainly driven primarily by personal beliefs. Exhibit 2-2 also shows that the 2014 Pertussis outbreak resulted in a decrease of exemptions from 1,080 in 2013 to 811 in 2014.



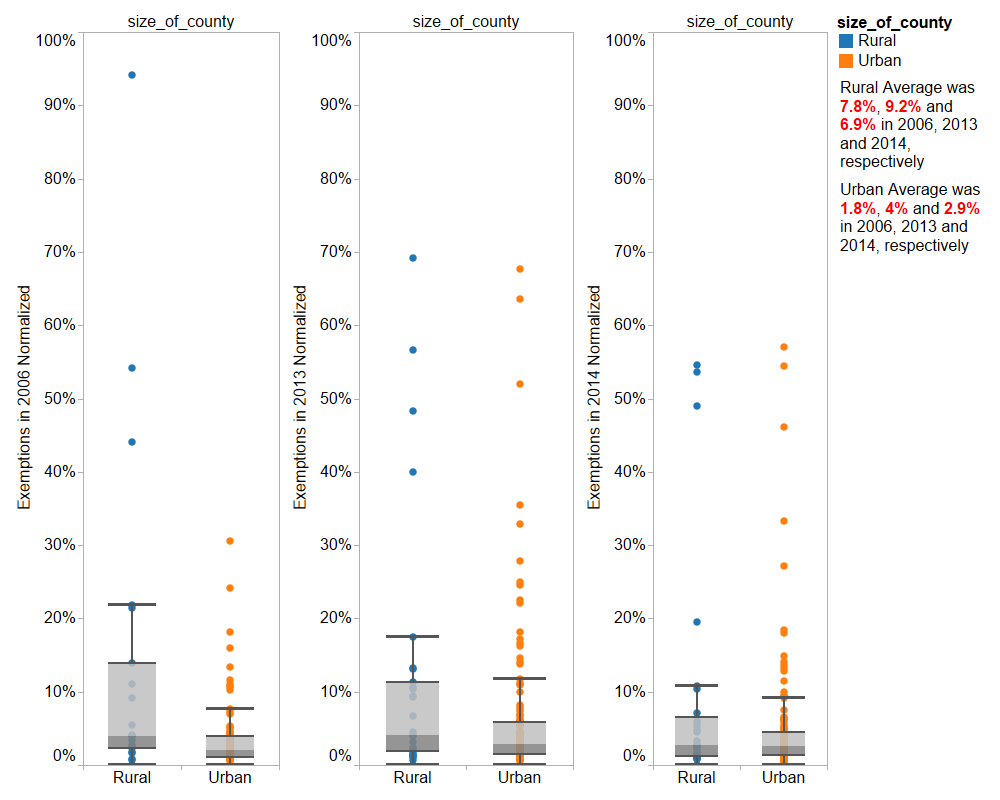
### Exhibit 2-2: Number of exemptions in the state of California from 2000 to 2015.

There were 37 different counties in the dataset used. Counties were classified as rural or urban to analyze if the size of the counties affected the number of exemptions. The Rural County Representatives of California (RCRC) is an organization that champions policies on behalf of California's rural counties. Their website has a list of rural counties that are members, which helped in identifying rural counties. Exhibit 2-3 shows that there are a higher number of exemptions in urban counties; this is due to the higher population and thus a higher number of students in urban areas. The other finding was that urban or rural counties had remained proportionally the same to each other for the three years selected.



### Exhibit 2-3: Number of exemptions by county size in 2006, 2013, and 2014.

Exhibit 2-4 shows the calculated percentage of exemptions at each school to normalize this metric. This report addresses the named percentage of exemptions and normalized exemptions interchangeably. This normalized number of exemptions was visualized in Exhibit 2-4 as three box plots to see the distribution of data alongside the average of the normalized number of exemptions in 2006, 2013, and 2014. The averages show that rural counties have higher amounts of exemptions than urban counties during the three years selected.



### Exhibit 2-4: Box plots of normalized number of exemptions during 2006, 2013, and 2014.

The calculated average percentage of exemptions enables us to compare rural and urban counties more effectively. The next step in the analysis was to figure out which counties had a higher average percentage of exemptions to observe if those also had higher Pertussis case rates. Plotting the average percentage of exemptions of all counties from 2006 to 2015 reveals that some counties such Riverside or Orange have a stable average percentage of exemptions. Exhibit 2-5 shows the counties with a stable average percentage of exemptions, and Exhibit 2-6 shows the counties with a non-stable average percentage of exemptions. Since the counties in Exhibit 2-6 like Napa show an increasing average percentage of exemptions, it will be easier to discover the effect of exemptions on case rates using counties like Napa from Exhibit 2-6.

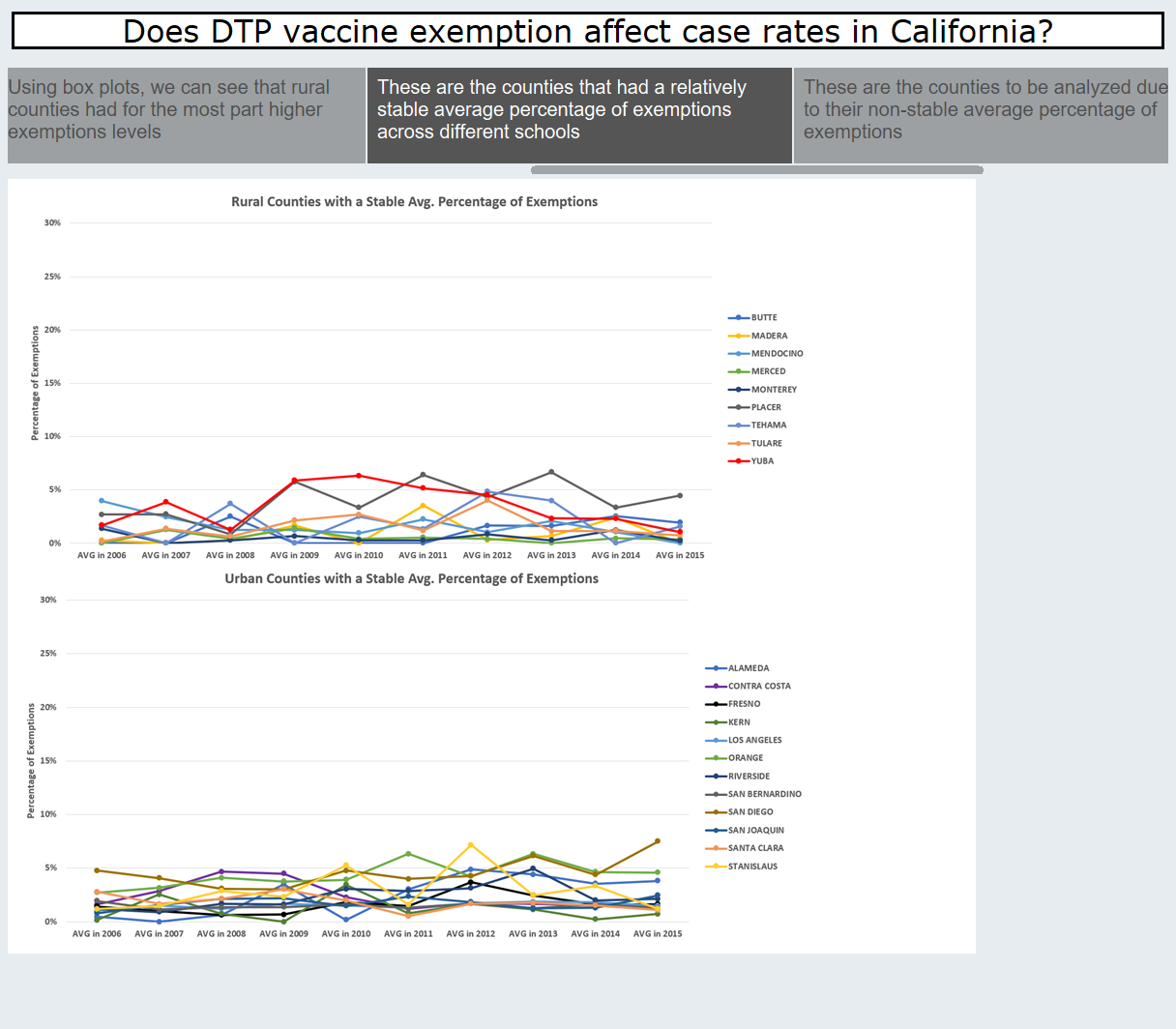


Exhibit 2-5: Counties with a stable average percentage of exemptions from 2006 to 2015.

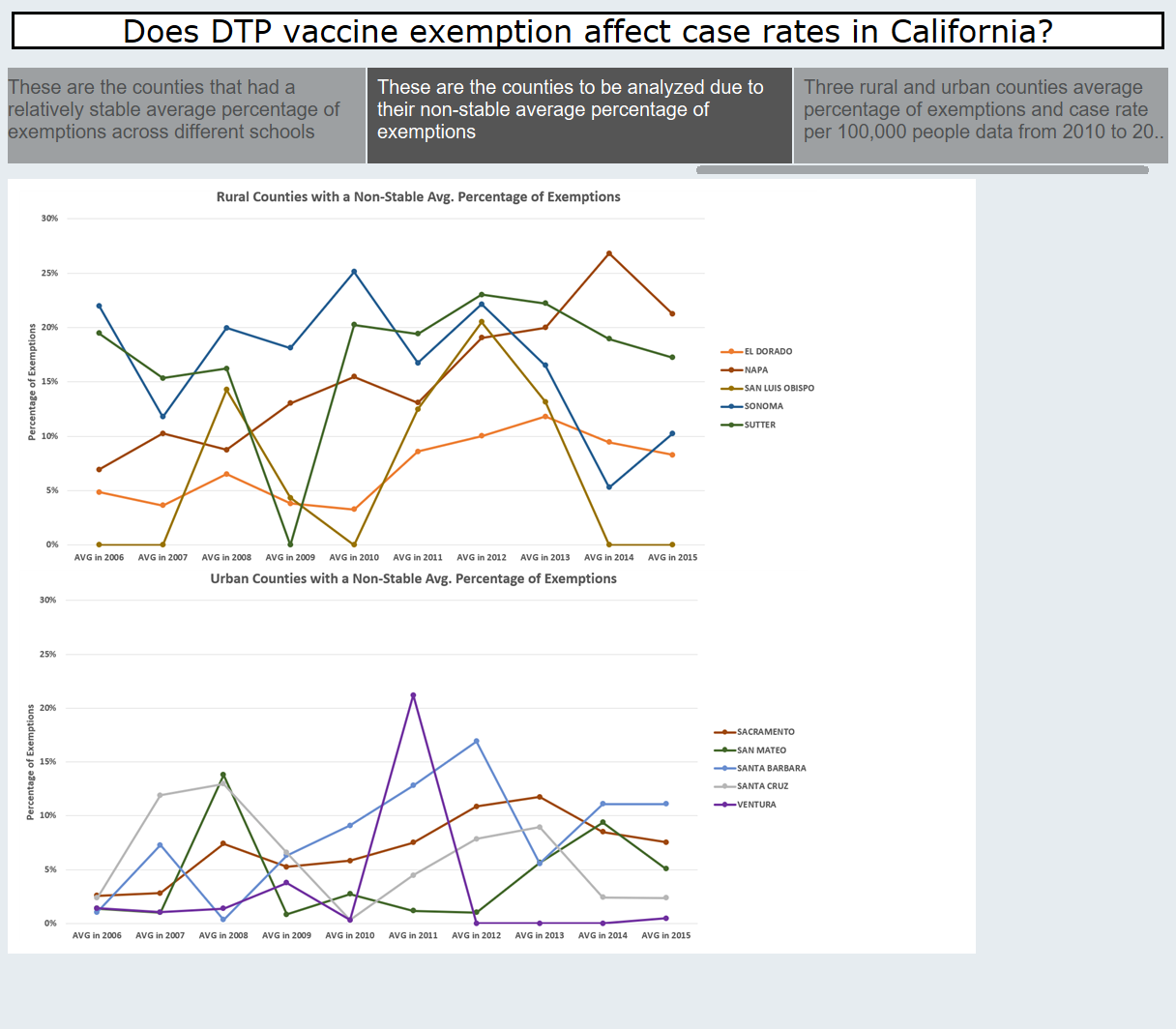
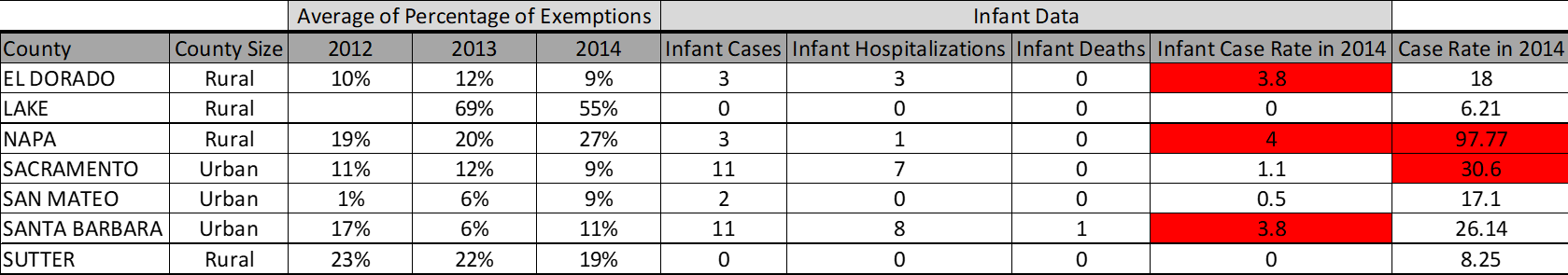


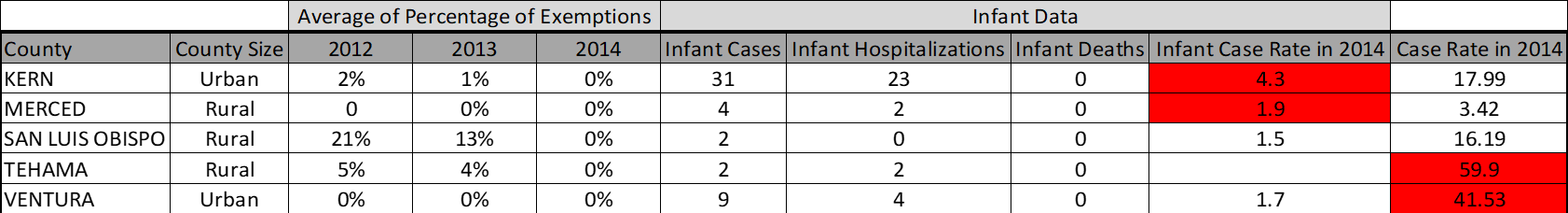
Exhibit 2-6: Counties with a non-stable average percentage of exemptions from 2006 to 2015.

Exhibit 2-7 shows the selected counties with the highest average percentage exemptions in 2014, which happen to also be plotted in Exhibit 2-6. The state of California average Pertussis case rate per 100,000 people in 2014 was 28. From the infant dataset, we know that the state of California average Pertussis case rate for infants in 2014 per 1,000 people was 1.8. Napa was the county with the second-highest average percentage of exemptions in 2014 at 27%. Napa was one of few counties showing a positive linear trend for the average percentage of exemptions during the past eight years. This county had an infant case rate of 4 (2.2 higher than the CA average during the same year) and a case rate of 97.7 (70 higher than the CA average during the same year). It was the only county shown in Exhibit 2-7 to have higher case rates higher than the California average for both infant and county-level case rates. Of the seven counties in Exhibit 2-7, four counties (Napa, El Dorado, Sacramento, and Santa Barbara) had either infant case rates or county-level case rates above the California average, as shown in red in Exhibit 2-7. The three other counties were Lake (highest average percentage of exemptions), San Mateo (Third-highest average percentage of exemptions), and Sutter (Seventh-highest average percentage of exemptions). Both Lake and Sutter had zero cases, hospitalizations, and deaths in 2014, which could be instances of missing data. Exhibit 2-7 shows that the majority of counties with a high average percentage of exemptions had either a higher infant case rate or county-level case rate when compared to the California average.

.

### Exhibit 2-7: Graph of the top seven CA counties with the highest average percentage exemptions in 2014. Case rates in red are higher than the CA average during that same year.

Exhibit 2-7 could prove that higher exemptions lead to a wider spread of diseases. However, the data also shows that there are other factors affecting case rates. For instance, there were five counties (Kern, Merced, San Luis Obispo, Tehama, and Ventura) with a zero-average percentage of exemptions in 2014, as shown in Exhibit 2-8. These counties also had case rates above the California average.



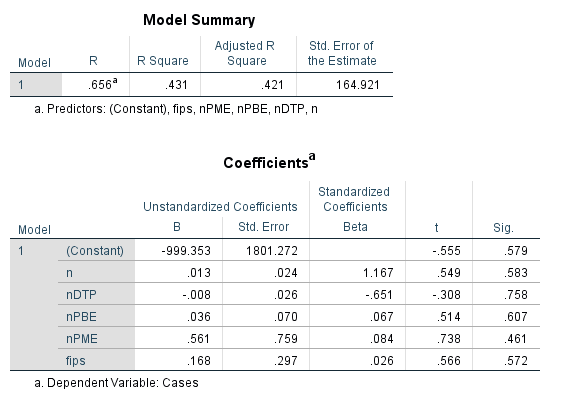
### Exhibit 2-8 - Graph of CA counties with zero average percentage exemptions in 2014. Case rates in red are higher than the CA average during that same year.

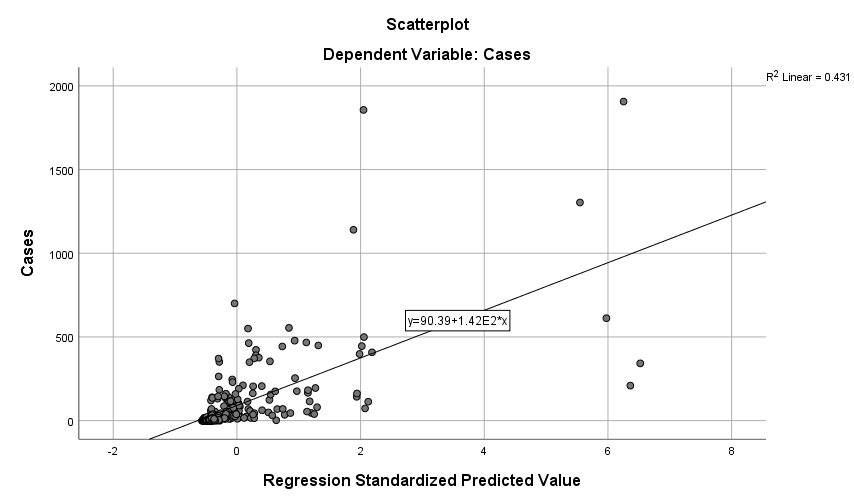
The data used to analyze the effect exemptions have on Pertussis case rates shows that having low exemptions within a county is not a guaranteed method to avoid high case rates. Vaccination is one of the multiple ways to manage diseases. The county of Napa is an example of a region where vaccination was certainly not a priority due to having an average percentage of exemptions increase from 7% in 2006 to 27% in 2014. This county had appalling infant case rates and county-level case rates during the Pertussis outbreak of 2014, which is a reminder of the importance of vaccination in schools.

## Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?

When attempting to answer the question of which California counties and schools had the most significant risk of facing a Pertussis outbreak, explorations revealed that the data did not track which school the Pertussis cases originated. Thus, pursuing which schools were at the highest risk, and by extension determining if the institution was private or public as a factor, was no longer feasible. Despite this setback, analysis by linear regression could still ascertain which elements leave a county at an elevated risk for a Pertussis epidemic.

IBM's SPSS software applied regression analysis using the number of cases of Pertussis among kindergarten students (Cases) as the dependent variable. Factors thought to be relevant to Pertussis case numbers served as the independent variables. These include the number of students in each county (n), the number of students who have received a Diphtheria, Tetanus, Pertussis vaccine (nDTP), the number of students with a permanent belief exemption (nPBE), and the number of students with a permanent medical exemption (nPME). Furthermore, to observe the potential effect of the counties themselves on case numbers, each county was represented using their unique numeric identifier from the Federal Information Processing Standard (FIPS).





### Exhibit 3-1: SPSS results from the initial model using all potentially relevant variables.

As shown in Exhibit 3-1, the initial model could benefit from some significant improvements; the initial R-squared value is quite low, and all independent variables show low significance ratings. These elements result in a poorly fitted model. As a remedy to improve upon this first model, the second model will employ quadratic regression to create additional variables and create a better fitting model.

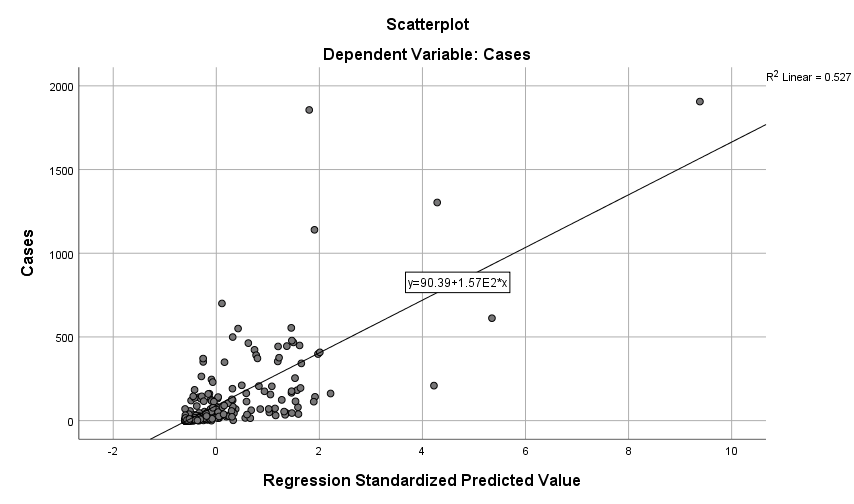
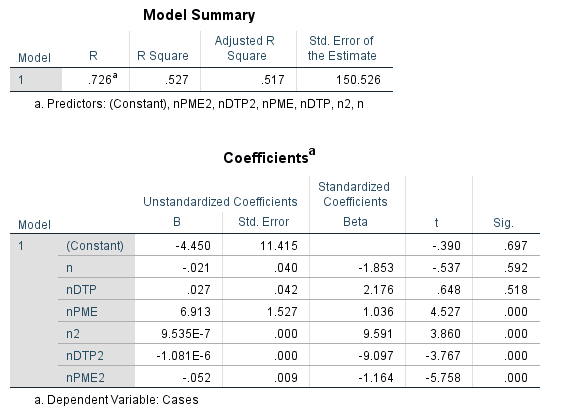


Exhibit 3-2: SPSS results from the improved model using quadratic regression.

Supplementing the squared predictors produces a moderate improvement over the initial model, incrementing the second model's R-squared value from .431 to .527. The FIPS and nPBE variables were found not to be relevant to this model, as they displayed very little significance in the previous model. Because of the low R-squared value, one may assume that there are other significant factors in determining the number of Pertussis cases a particular county faces. Based on the SPSS output, 52.7% of the variance in the number of cases can be explained by the following equation:

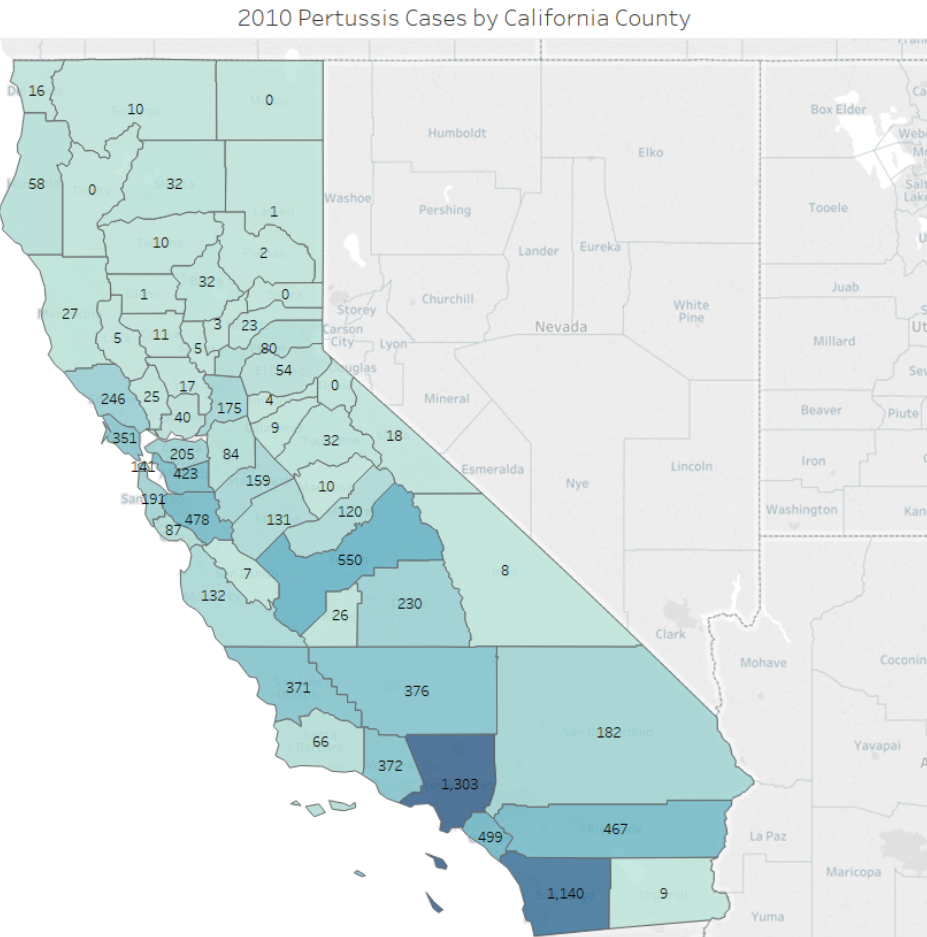
*Cases = nx1 + nDTPx2 + nPMEx3 + n2x4 + nDTP2x5 + nPME2x6*

The equation, as defined above, is applied in R Studio and initialized to create regression trees to detect which variables were the most influential in determining the number of cases and serves as a visual representation of the multiple linear regression.

### Exhibit 3-3: Regression Tree for determining the number of cases.

### Exhibit 3-4: Minimum Error Regression Tree for determining the number of cases.

From examining both the standard regression tree and the minimum error regression tree, it is clear that n (the number of kindergarten students in each county) is the primary factor in determining the number of cases. The analysis reveals that counties with more than 43,000 students have the highest potential risk of a Pertussis outbreak, specifically Los Angeles and San Diego. Findings also convey that counties with greater than 14 students claiming permanent medical exemptions and having more than 34,000 students are also at an elevated risk of a Pertussis outbreak, namely Orange County and particular years in Riverside.



### Exhibit 3-5: Map of 2010 Pertussis cases among kindergarten students by California county.

Tableau is employed to map the 2010 Pertussis case data against the findings. 2010 marked a significant Pertussis outbreak, with more than 9,000 cases reported in the state. As expected, Los Angeles County and San Diego County exhibit by far the most documented number of cases. Orange County and Riverside County echo the analysis with just under 500 cases each, putting them amongst the counties with the highest number of cases.

Several outliers and anomalies appeared during the analysis. Noticeably, Fresno and Santa Clara County with 550 and 478 cases, respectively. Both counties, Fresno, in particular, have much lower values of n than other counties that feature a fewer number of recorded cases. These outliers likely come as a result of the chosen regression model's low R-squared value of .527. Thus, it is implied that there are elements not found in the data sets that may be of significant value in determining the potential number of cases.

# Conclusion

## Question 1: Does a location’s economic situation affect the rate of immunization among school children?

It is evident that no single factor can be definitively linked as the determining factor of a California county’s immunization rate among its school children. While a location’s economic situation cannot be excluded as a point of influence, it is not the defining determinant. There appears to be a multitude of variables affecting the ultimate decision of whether a child will be immunized. Take, for instance, household income and immunization for two very different counties of California - San Mateo County and Lake County. San Mateo County’s median family income is over six figures while Lake County’s is $42,475 - below both the national and state averages[[1]](#footnote-1). The aggregated immunization rate measure calculated for Question 1 is actually higher for Lake County at 92.4% compared to San Mateo County which has a rate of 91.4%. Similarly, counterintuitive examples can be found when comparing California’s counties: rich and impoverished; rural and urban; highly educated and those less so.

If research on this question is to be furthered, additional strategies should be considered. For instance, time series-based approaches can be used to see if the level of immunization increased or decreased over time, which can help a researcher isolate potential variables by comparing with other factors over the same period. Qualitative methods, such as interviews, may also prove useful in gaining insights as to why some families choose not to immunize. The use of simple linear regressions should not be discounted if data at a more granular level can be obtained since relationships may become more apparent at the city/township or zip code level. In addition, it may also be worth factoring in other variables such as household size, level of health insurance coverage, distance to hospitals/clinics, and whether the parents were immunized. Further research is warranted as public health authorities would be better informed and ready to intervene before a region’s herd immunity is compromised.

## Question 2: Does DTP vaccine exemption affect case rates in California?

The state of California has not been immune to the effects of Anti-Vaccination movements that foment the “negative side effects” of vaccines. The number of vaccine exemptions in California schools has increased rapidly during the past 10 years. Since contagious diseases such as Pertussis are cyclical, a family’s decision regarding vaccination is highly dependent on how active diseases are. After outbreaks, some families do prefer vaccination putting aside any Anti-Vaccination beliefs. There are many counties in California, both rural and urban, with a stable average percentage of exemptions. These counties had relatively low levels of exemptions for many years. Sadly, they were still affected during outbreaks. There were also counties with an increasing average percentage of exemptions like Napa with appalling case rates that were much higher than the California average case rate. During an outbreak, having a low number of vaccination exemptions should not be the only method to manage diseases. There is no evidence that counties with low exemptions were 100% immune to a fast-spreading disease. Nonetheless, if the number of exemptions is not controlled like in the case of Napa, it gives a disease the chance to cause more casualties.

## Question 3: Which California counties and schools have the greatest risk of facing a Pertussis outbreak?

Higher density California counties and schools have the most significant risk of facing a Pertussis outbreak. Analyzing the number of kindergarten students per county reveals a consequential effect in counties with 43,000 or more students.

While the number of kindergarten students in each county is the primary factor in determining the number of cases, several outliers were apparent. Fresno and Santa Clara each displayed a higher case count than other counties with fewer overall kindergarten students. However, these outliers are more likely due to the chosen model's relatively low R-squared value. It implies that there are elements not found in the data sets that may be of significant value in determining the potential number of cases. Counties with reports of permanent medical exemptions greater than 14 students reduce the student population requirements considerably, as demonstrated in Orange and Riverside county.

# Appendices

## Appendix A: SQL Code

#### SQL query to combine and create CountyRates table, listing each element as a sum by year and county.

Create Table CountyRates

AS

SELECT Public."rates"."county", SUM("n") as "n", SUM("nDTP") as "nDTP", SUM("nPolio") as "nPolio", SUM("nPBE") as "nPBE", SUM("nPME") as "nPME", "Cases", "Rates", Public."rates"."year"

FROM Public."StudentData", Public."rates"

Where rates."year" = Public."StudentData"."year" And rates."county" = Public."StudentData"."COUNTY"

Group by "county", "Cases", "Rates", Public."rates"."year"

Order by "county", Public."rates"."year"

## 

## 

## Appendix B: R Code

R script used to group and aggregate 2010 California Kindergarten Immunization Data into county level measures.

StudentData\_2010 <- StudentData\_2010 %>%

group\_by(COUNTY) %>%

summarise\_at(vars(n, nMMR, nDTP, nPolio), sum)

StudentData\_2010 <- StudentData\_2010 %>%

mutate(score = rowSums(.[3:5]))

StudentData\_2010 <- transform(

StudentData\_2010,

score = score / n / vaccine\_count

)

#### R script used to create a table to average percentage exemptions for all California counties

AVG\_EXEMP\_NORM\_COUNTIES <- unique(student\_dat[,3:4])

colnames(AVG\_EXEMP\_NORM\_COUNTIES)<- c("COUNTY","size\_of\_county")

for (i in 2006:2015){

##obtain data from 2006 to 2015 and save to student\_dat\_loop

student\_dat\_loop <- student\_dat[which(student\_dat$year==i),]

##find the means

agg\_mean\_per\_county\_loop <- aggregate(student\_dat\_loop$total\_exemptions\_normalized,by=list(student\_dat\_loop$COUNTY),FUN=mean)

colnames(agg\_mean\_per\_county\_loop) <- c("COUNTY", sprintf("AVG in %d",i))

##Merge all the averages from 2006 to 2015

AVG\_EXEMP\_NORM\_COUNTIES <- merge(x=AVG\_EXEMP\_NORM\_COUNTIES,y=agg\_mean\_per\_county\_loop,by="COUNTY",all=TRUE)

}

##Export file

write.csv(AVG\_EXEMP\_NORM\_COUNTIES, "AVG\_EXEMP\_NORM\_COUNTIES\_FROM\_R.csv", na = "", row.names = FALSE)

#### Script used for creating regression trees.

setwd("E:/Users/Colin/Documents/Homework/ISDS 577")

data = read.csv('CountyPertussisData.csv', stringsAsFactors=T, head=T)

library(rpart)

library(rpart.plot)

set.seed(1)

fit = rpart(Cases ~ n + n2 + nDTP + nDTP2 + nPME + nPME2, method = "anova", data = data, cp = 2e-3)

rpart.plot(fit, main = 'Regression Tree') #Standard Regression Tree

pfit = prune(fit, cp = fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])

rpart.plot(pfit, main = 'Minimum Error Regression Tree') #Minimum Error Tree

1. Based on US Census Bureau QuickFacts 2014-18 (in 2018 dollars) [↑](#footnote-ref-1)