Gagnavinnsla

Hópverkefni

Benedikt Þorri Þórarinsson
Grétar Ingi Guðmundsson
Hinrik Helgason
Kristján Ingólfsson

Kennari: Eyjólfur Ingi Ásgeirsson

Introduction

For our project we decided to use several datasets full of information about vaccination rates, population numbers and other useful statistics on California counties. The purpose of the project was to answer several questions about a number of variables such as population, population density, gdp per capita and political affiliation and their relation to the very high rate of pertussis (colloquially known as whooping cough) that plagued the state in 2014.
Finding connections like these can be an important tool for maintaining good public health and shaping policy.

The Questions

What we were perhaps most interested in finding out was if there were some sort of connection between political affiliation and the prevalence of pertussis among the population of a county. The stereotypical anti-vaxxer is often described as either an urban liberal who only consumes natural and organic products or a more rural and highly conservative government sceptic. We set out to see if there would indeed be a liberal/conservative divide or not.
The relationship between gross domestic product per capita and the rate of pertussis was also something we wanted to explore. Were wealthier counties more or less likely to vaccinate and/or have a higher rate of the disease? Population density was another variable that we examined. Would the more densely populated counties have a higher rate of the disease then the less densely populated ones?

Methods

Deciding the theme of the project

Initially we had chosen cars as a topic and how and why certain types of cars were more popular in countries like Russia or Germany. However, we had trouble finding datasets with data about car sales in those countries so we decided to change our theme entirely. Instead we chose to investigate if people who preferred to not have their children vaccinated were more likely to elect republican or a democrat in the 2016 presidential elections, i.e. Donald Trump and Hillary Clinton. This was mainly decided as we found sufficient data for this topic, especially for pertussis which is the disease we focused on the most. We also found it quite interesting as vaccination has been a controversial topic in the past few years.

Finding datasets

The datasets that we had at our disposal were discovered on various websites. On *kaggle.com* we found a long list of elementary schools in California, sorted by county, and the prevalence of pertussis vaccination among the students in each year from 2000 to 2015. Not every one of these years is represented in every school, but it is still an invaluable resource.

In addition to this huge list we had a comprehensive list of all votes cast during the 2016 presidential election which was discovered on *dataworld.com*. This list was sorted by county and/or the town where the ballot was cast. From this list we gathered the data about the counties in California and used them to determine the political leanings of the inhabitants of each county.

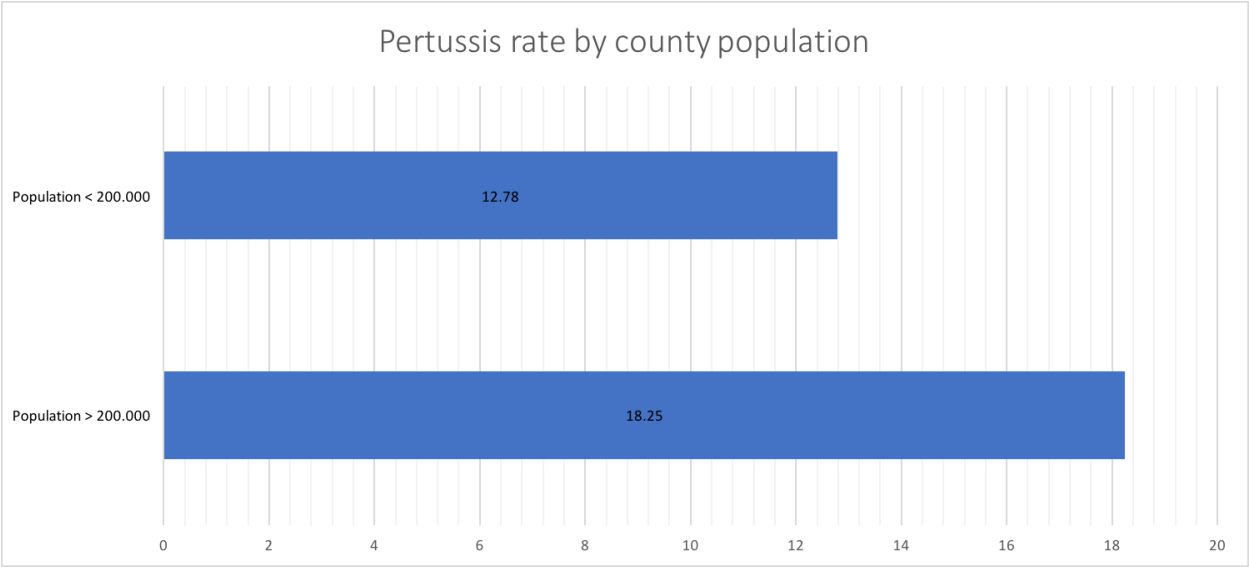
Finally, we had two smaller tables which contained vital information on each county in the state. One had population numbers, population density, gdp per capita and more and another had the rates of pertussis in each county from 2010 to 2014. The data in those tables was found on *wikipedia.com* which was converted to a .csv format.

Creating tables for the database

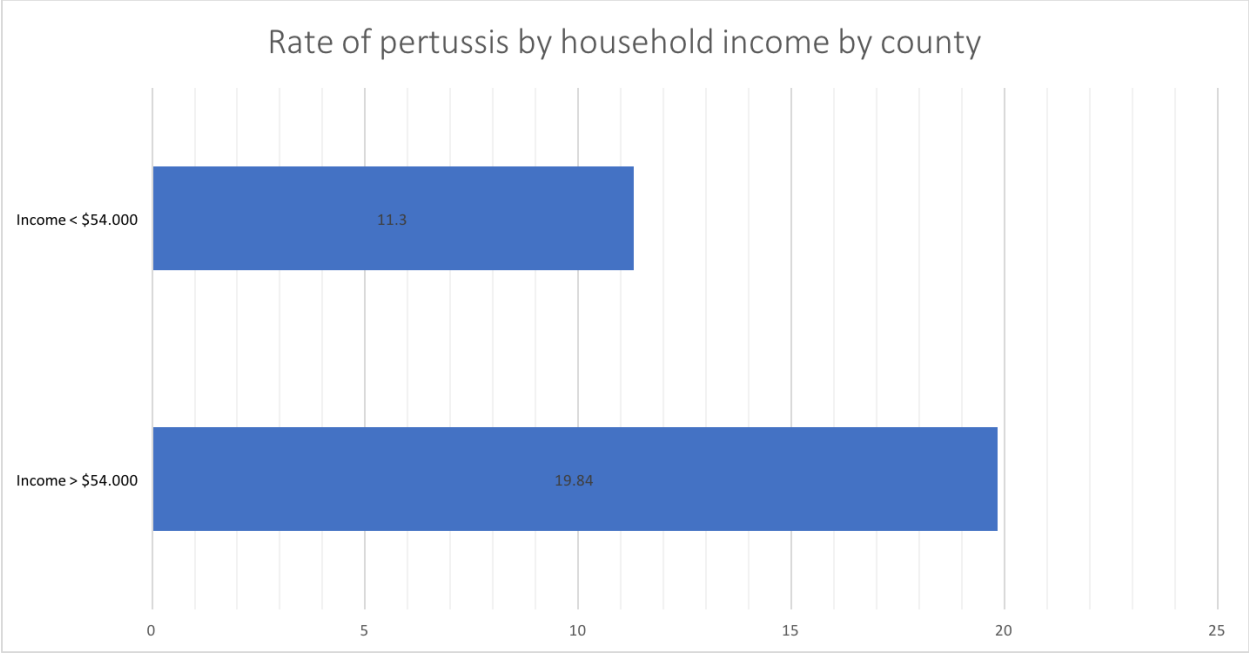
Using the datasets we had at our disposal, we first created a table scheme which changed a bit with time as we discovered new questions to answer and/or we thought that our current database lacked data when we started coding.

When the tables had been created, we wrote a python file that would create the SQL insert commands which we'd use to insert the data into the database. At first, the data had to be inserted manually via the copy-paste method into *datagrip*, which is a development environment. This method proved to take too long as the database contained over 100,000 rows of data so the rest of group members managed to find a shortcut to get the data on their computers.

Results

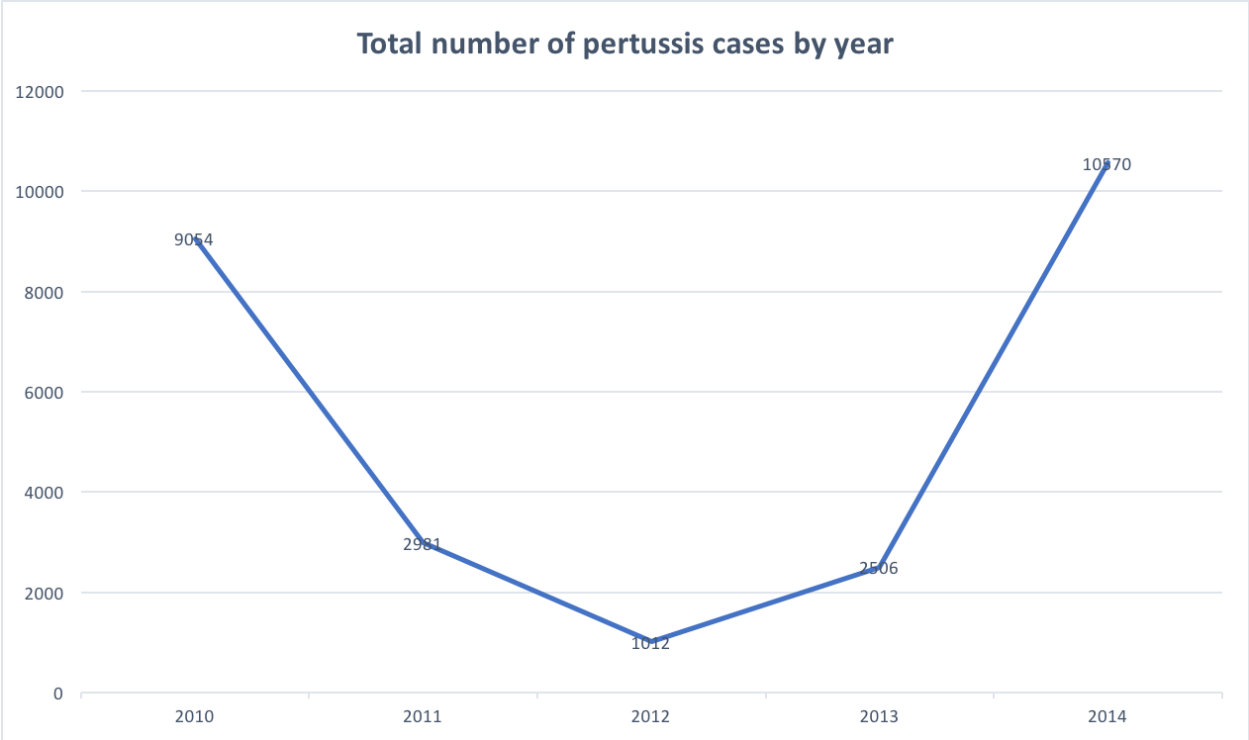


This graph shows counties with population above 200.000 have much higher rate of pertussis than the counties that have population below 200.000. It is well known that diseases spread more easily in dense areas like cities than in less dense areas.

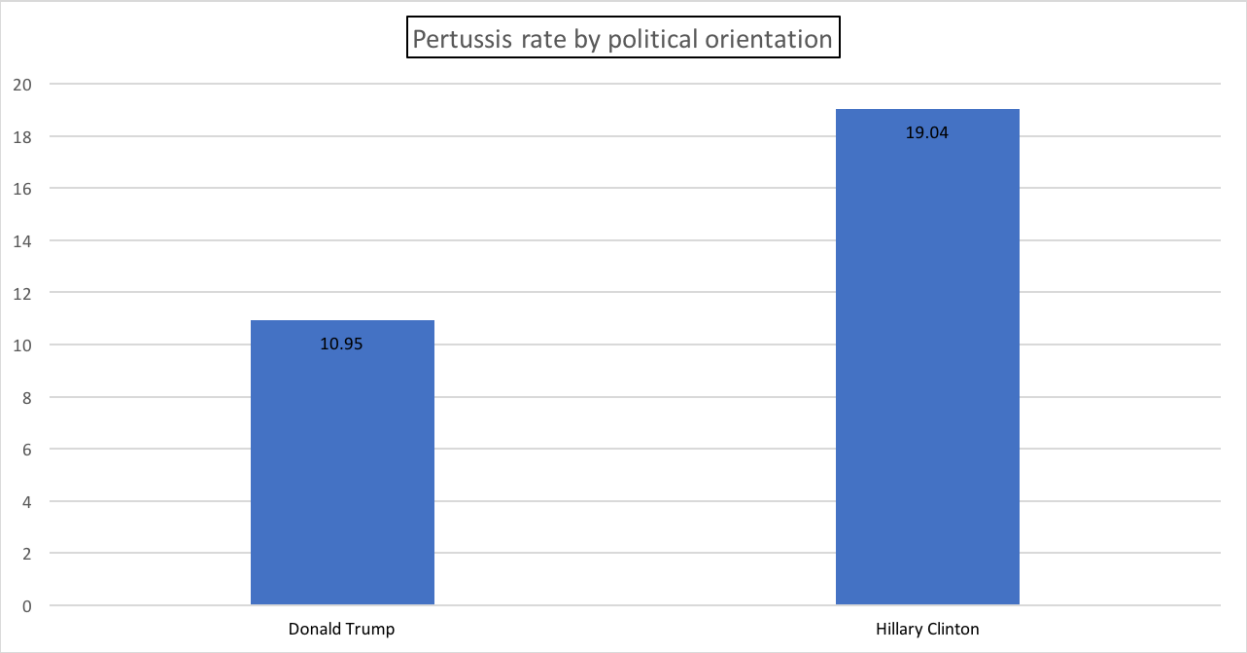


This graph shows that counties with average household income above $54.000 get almost twice as much pertussis cases than counties with the average household income below $54.000. It is a fact that people who live in the cities have higher salary than those who live outside the city so

it is normal that there are more cases of pertussis by counties that have more income because they also have higher population density.



This graph shows the number of pertussis cases per year in the state of California between 2010 and 2014. The disease, as any other, goes through cycles with some years being more severe than others. An interesting fact is that since the 1960s and 70s the disease had been almost completely eradicated through vaccination, but in the 80s and 90s there was a huge increase in the number of cases. Finally, in 2010 there was a record number of cases, almost certainly caused by the wave of anti-vaccination sentiment that started in the 80s and 90s.



This graph represents the average number of people affected by pertussis based on political orientation(or rather based on the political orientation of the counties as a whole). What it's telling us that if an individual voted for Hillary Clinton, it's far more likely that their child will get pertussis than if he/she voted for Donald trump.

Discussion

Our finding suggest that the voters of Hillary Clinton are far more likely to have pertussis than those that voted for Donald Trump. This is based on the average pertussis rate of each county from 2010 to 2014 compared with the voting data for each county in 2016.
At first glance this might suggest that liberal voters are less likely to vaccinate their children, but as it turns out the rates were quite similar to those of Trump's counties. Lower that the 95% rate suggested by the World Health Organization as the rate required for fully effective herd immunity against the disease, but still usually around a decent 80-90 percent.
Another factor that we suspected was the population density. It is well known that cities are far more vulnerable to epidemic diseases than rural towns and villages. We didn't have any data to back up this hypothesis so it would have to be the subject of further study.