Yikai Zhang

UCLA Extension

Fundamentals of Data Science

**Final Report**

Inspiration: December 1 is World AIDS Day, it is a day dedicated to raise awareness of the AIDS caused by the spread of HIV infection. In mourning those who have died of the disease and help people to get more knowledge of HIV disease, I plan to do a term project on HIV.

Question: What age group in state of California has the largest number of people affected by HIV? Which spreading method is most effective? Is the trend going to continue in the future?

Dataset: due to the data scarcity on HIV transmitted population around the world and time limitation, the dataset for state of California is a better option for this project. The data is from Healthdata.gov, below is the link of dataset:  
<https://healthdata.gov/dataset/hivaids-cases>

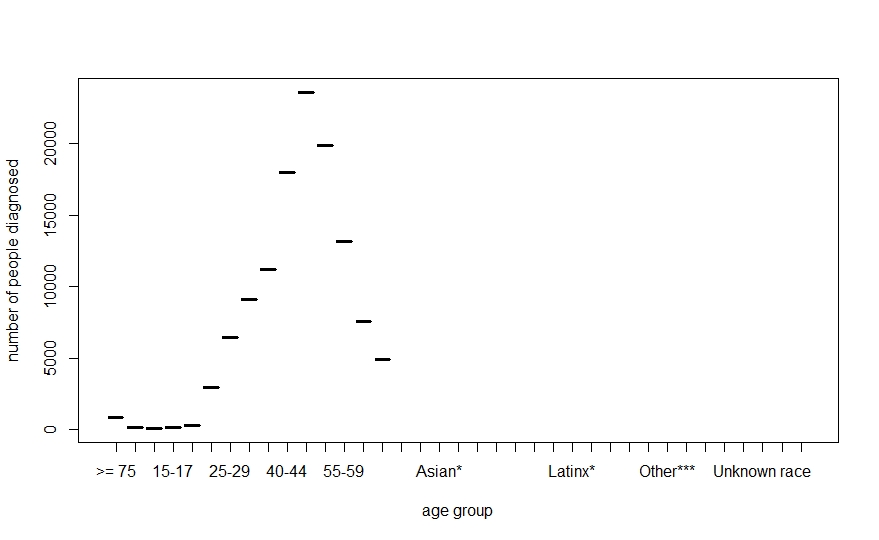
Data analysis:

The dataset for this project contains several categories, a well-organized data is a key to successful project, therefore I separate dataset into 4 smaller tables based on their category, they are age, gender, race and transmission respectively.

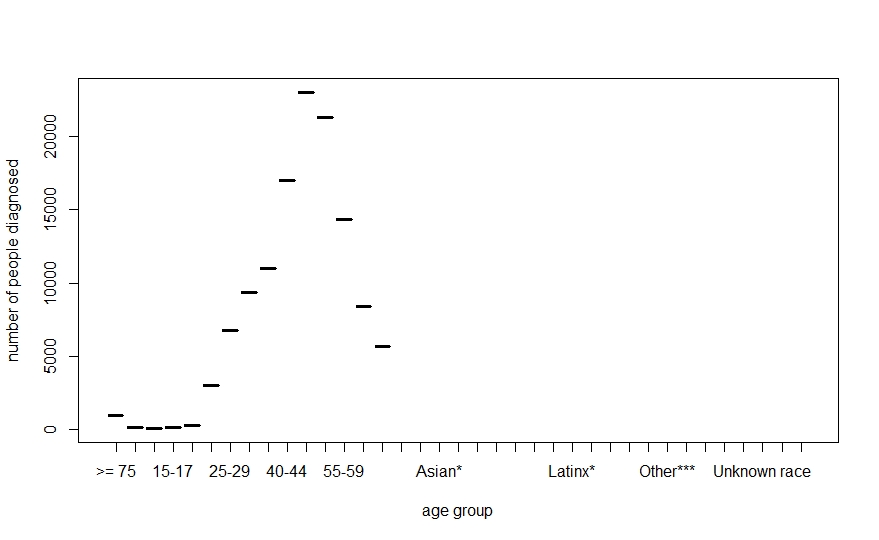
First, I calculate the total number of people that are diagnosed with HIV infection each year, and as you can see from the R script, the number has been increasing, from 118,487 in 2011 to 135,082 in 2017.

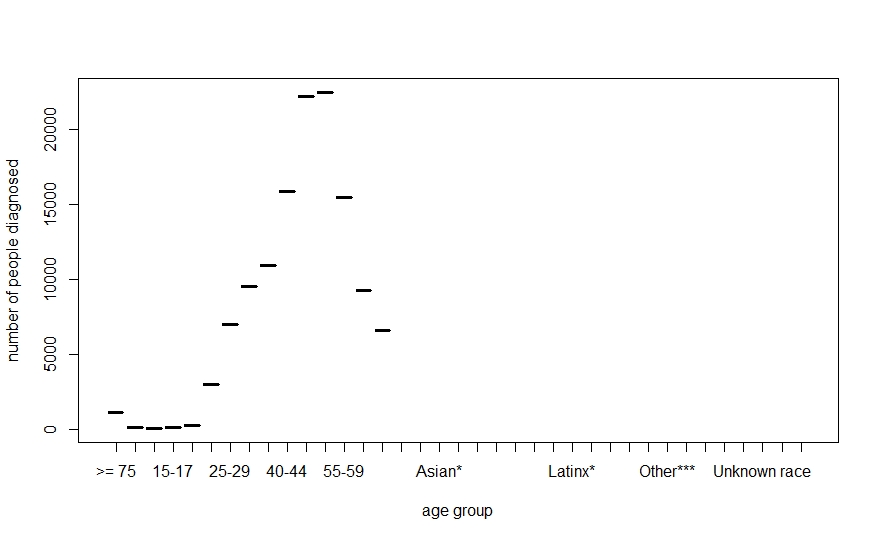
Then I plot population distribution for different age groups each year, below are the results.

2011:

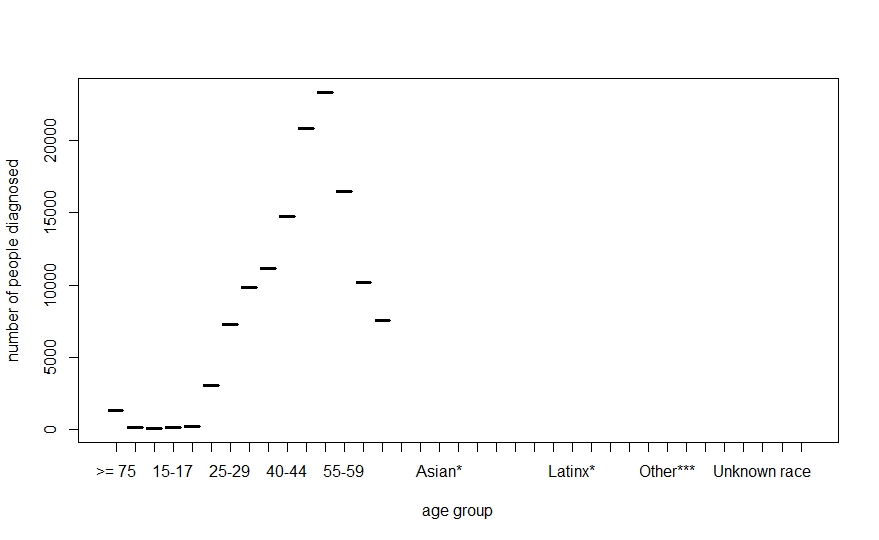


2012:

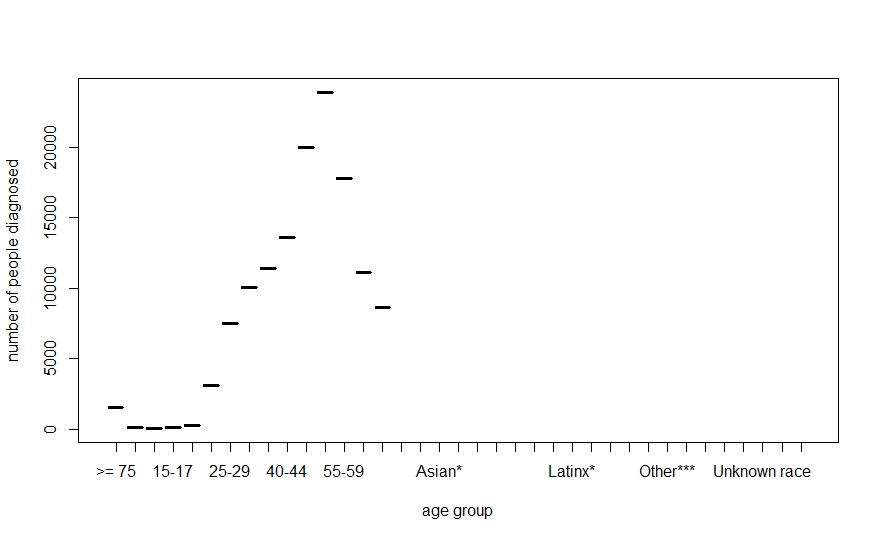


2013:

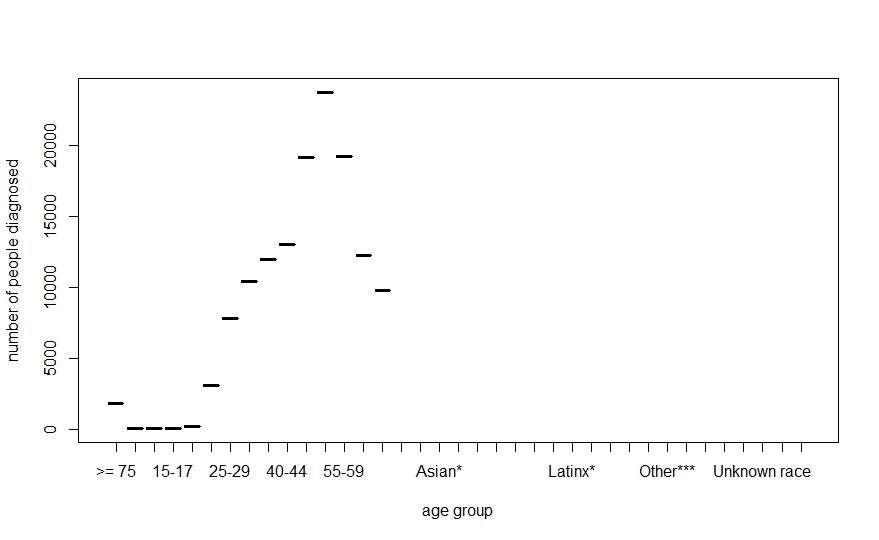
2014:



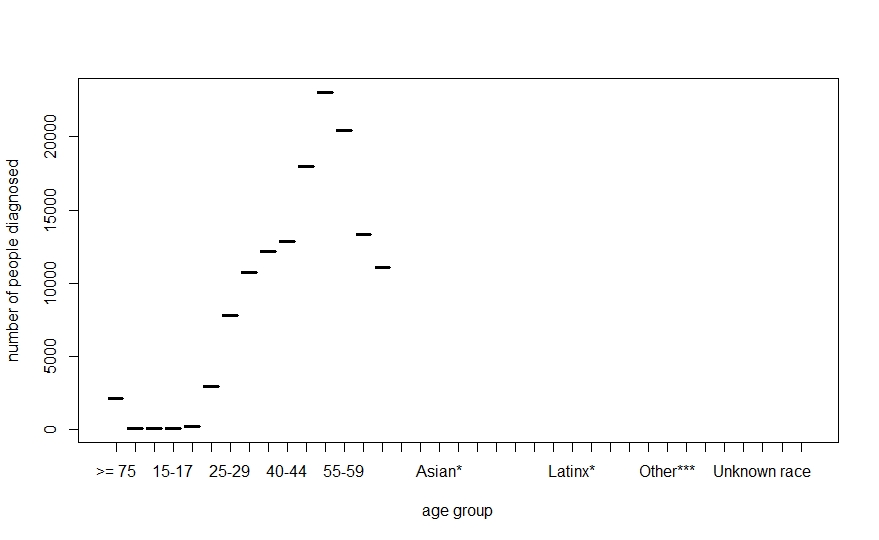
2015:

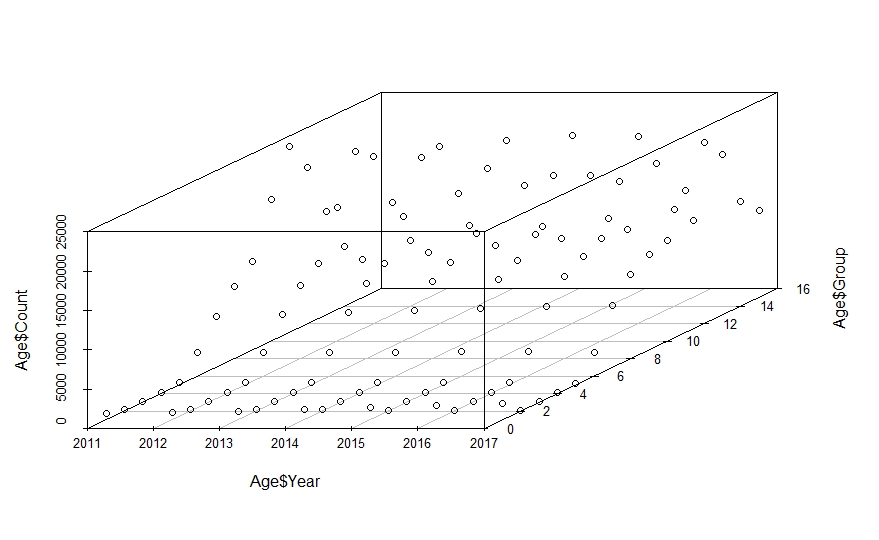


2016:

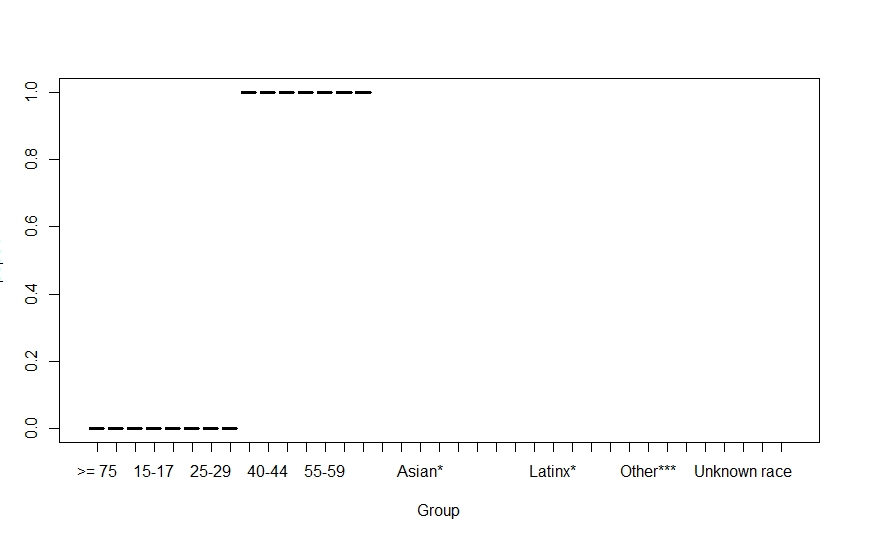


2017:

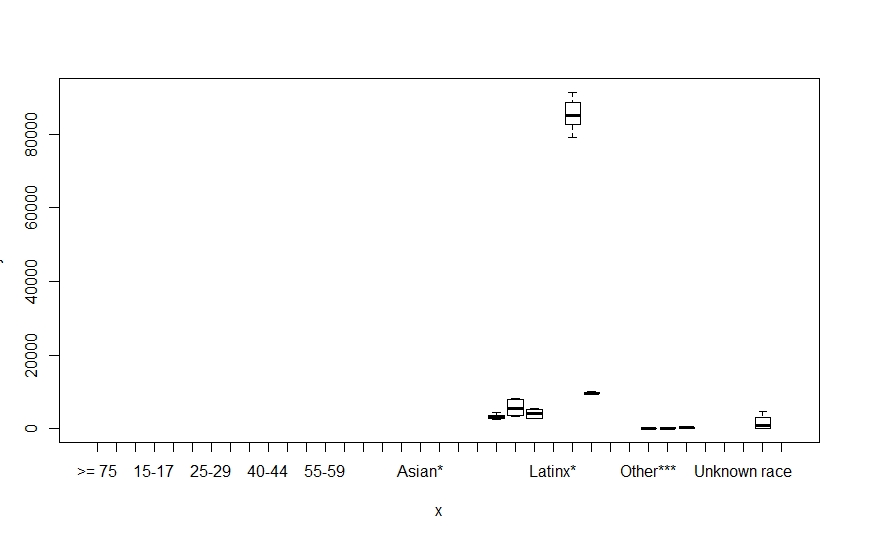




As you can see in the above plots, the number of people diagnosed with HIV infection is high among age group from 45 to 55 for all the years, and the last 3d vision image gives us the direct look on how numbers are distributed. From 2011 to 2017, the number of people diagnosed with HIV virus has been increasing for each age group, and we can predict that the trend is likely to continue. I also created a binary plot for Age data 2017, it explicitly states that among people are infected by HIV disease, people who are older than 40 years old consists the most part.



Now that we have observed data on HIV in age group, we can jump into analysis on which spreading method plays an important role in virus transmission.



The above image shows the relation between transmission type and number count. The graph is sketchy because on the x bar it shows other unrelated factors such as age and race. Since the transmission data was directly derived from hivdata1 table, it still keeps the format of original table. As the statistics show, the male-to-male sexual contact has the greatest number than any other groups.

Outcome: As I have illustrated in many plots, the number of people who are affected by HIV virus is increasing, and male-to-male sexual contact seems to be the major reason for HIV infection. The number of people who are infected by HIV virus will very likely to increase in the future as the death rate goes down. With the development of medical technology, the virus can be well controlled.

Memo: As my first ever data project, there are many shortages both in my report and script, and I could have chosen better algorithm methods for some questions. The current stage is only a small part of bigger project, and there are many more areas that I need to improve on for this project. I will keep revising on my script and outcomes, adding other dataset such as viral loads to make a better scope. Thank you for taking your time on reviewing it.