Extra Credit Assignment

**Introduction**

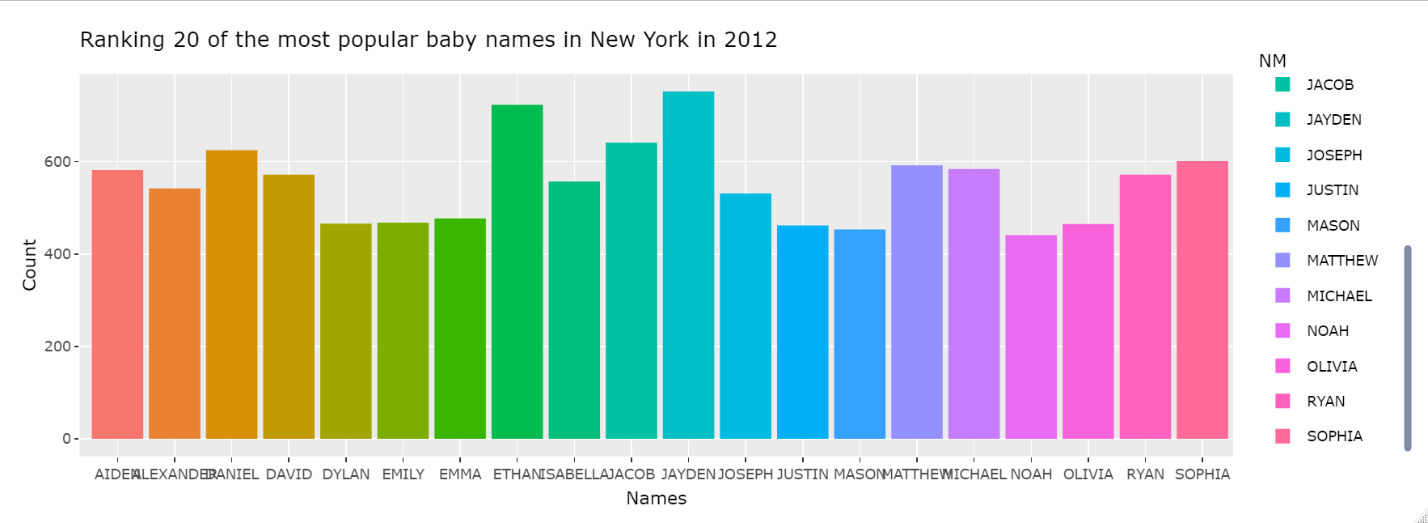
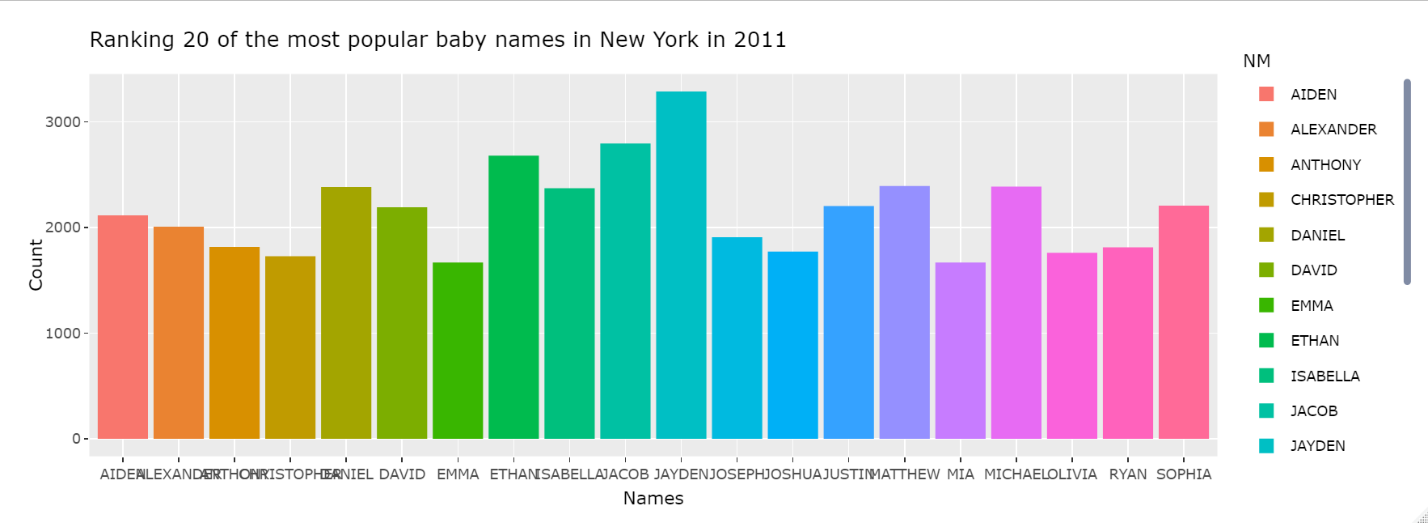
This data set measures the most popular baby names in New York between the years of 2011-2014. The data set was collected through the civil birth registration and was collected by the Department of Health and Mental Hygiene. The data was collected for the purpose of ranking the popularity of baby names specifically within New York City. The data set has a number of variables such as the baby’s names, their gender, their mothers’ ethnicity, the year they were born and the count of the name within that year and the ranking of that name.

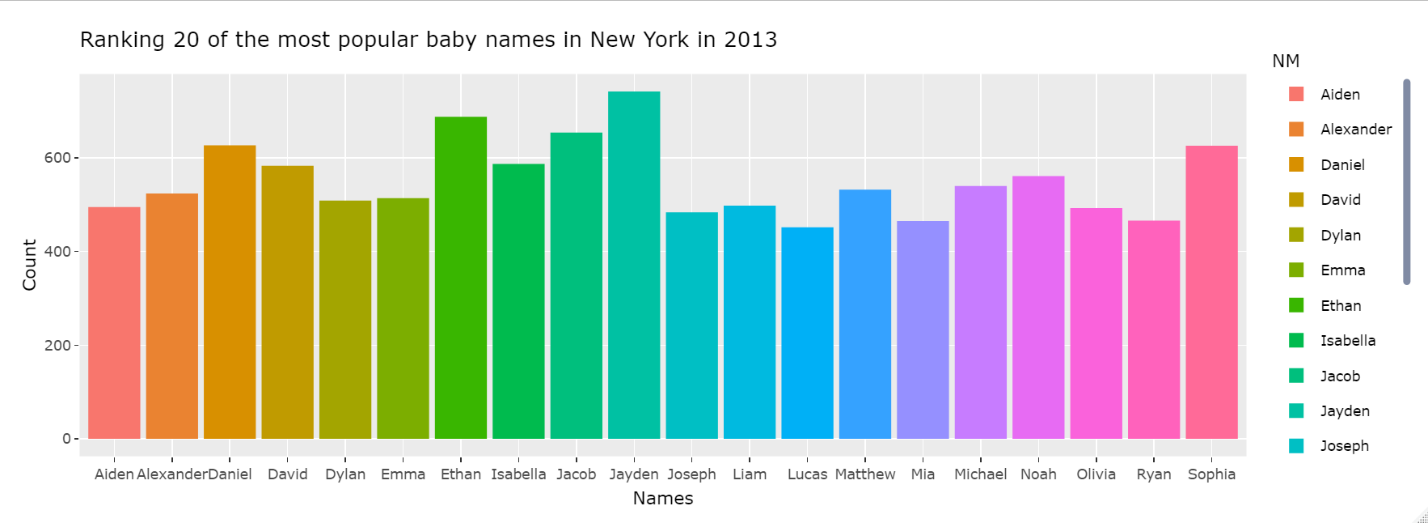
**Data Visualization**

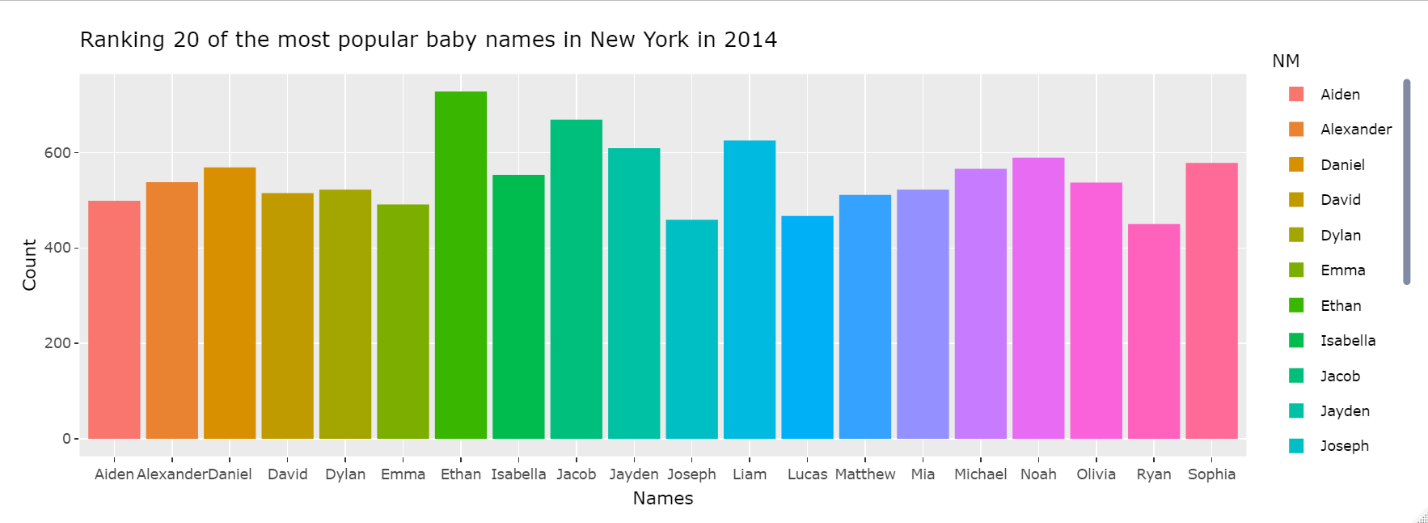
It’s an interactive data visualization that you can view in the R file.

<https://github.com/Jzorx/ExtraCredit1>

All the plots are at the bottom of the code to view in the names.R file.





****

**Discussion:**

In these visualizations we compare the most popular names in each year, from 2011-2014 to analyze how name popularity can change across each year. People that might be impacted by this data might be parents who would potentially want or not want their kid to have a common name. In my case, my parents wanted to give me an orthodox name. Another further discussion that can be had is maybe with further analysis and machine learning we can use this data to predict name popularity in future years.

**Reflection:**

In conclusion from my analysis I can conclude that the most common name across 3 of the 4 years was Jaden which surprised me. We can also see from the graphs that the popularity of the name Jayden decreased in 2014. Also, how the most common names are predominantly white/Caucasian names suggesting the influence of history still having an impact on names today. Regarding the data set however, one thing stood out to me which is also apparent on the graph. With 2011, the counts are a lot higher than 2012, 2013 and 2014 suggesting that either a lot more people were born in 2011 which is unlikely, or that the data set has a problem that there are a lot more entries for 2011 meaning that too many entries were added for 2011, or that there is data missing from people born in 2012, 2013 and 2014. The fact that 2012-2014 are consistent with each other leads me to believe that extra entries were added to 2011, however it is also more likely in general that data is missing rather than fake data being added.