Project Practice

Christy O'Brien

2023-08-30

Data Setup

The data was downloaded from the city of New York. The original researchers got this information for the civil birth registry. Extra information is provided at the end of this document.

```r
knitr::opts_chunk$set(echo = TRUE)

# Clear RStudio's memory
rm(list = ls())

# set your working directory
setwd("C:/RFiles/MCC")

library(tidyverse)

# https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf
names_data <- read_csv("Popular_Baby_Names.csv")
```

Cleaning Data

Some column names included spaces which made it inconvenient to code, so those columns were renamed. Additionally, some the names were coded with names in all caps while in other entries the same name had only to first letter capitalized. Since R is case sensitive, this may lead to invalid results. Thus, all names were made all uppercase to ensure that the same name was not split into different observations.

```r
# https://sparkbyexamples.com/r-programming/rename-column-in-r/
names_data <- names_data |>
  rename("year" = "Year of Birth", "name" = "Child's First Name")

# standardize name entry
names_data$name <- toupper(names_data$name)

# count how many ethnic names in a year per grouping
names_data_by_count <- names_data |>
  add_count(year, Ethnicity)

# make year an integer
names_data_by_count$year <- as.integer(names_data_by_count$year)
```

```
# standardize ethnicity entries
Asian <- c("ASIAN AND PACIFIC ISLANDER", "ASIAN AND PACI")
Black <- c("BLACK NON HISP", "BLACK NON HISPANIC")
white <- c("WHITE NON HISP", "WHITE NON HISPANIC")

ethnicities_blocked <- names_data_by_count |>
  mutate(Ethnicity = case_when(Ethnicity %in% Asian ~ "ASIAN",
                               Ethnicity %in% Black ~ "BLACK",
                               Ethnicity %in% white ~ "WHITE",
                                 TRUE ~ Ethnicity))

labels <- c(2012, 2015, 2017)
```

Plotting Ethnic Naming Trends

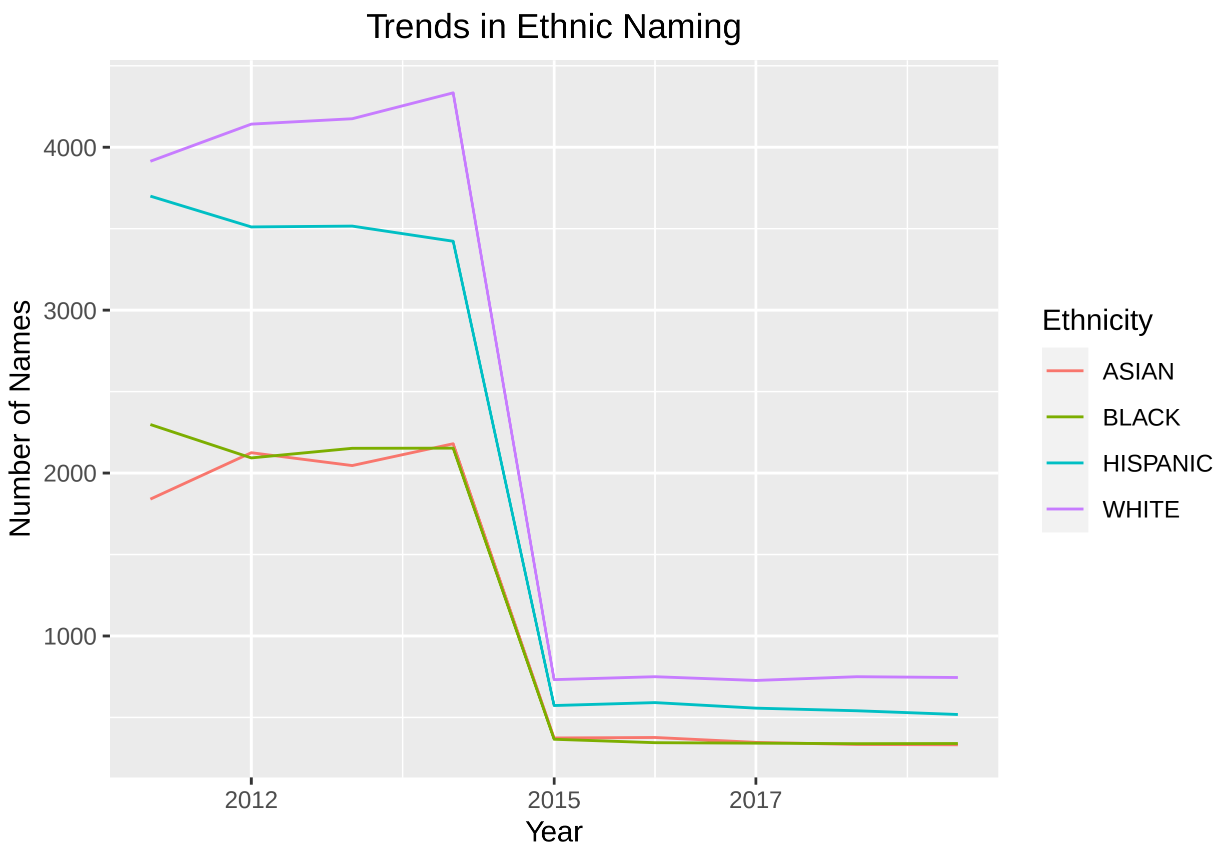
The first line graph groups names based on ethnic background and plots their prevalence through time. This should suggest whether certain naming conventions are increasing, decreasing, or remaining constant.

There is also a stacked bar graph which shows the proportions of how many children of each ethnicity are named the most popular names. A line graph with the most popular names is also provided to show how ethnic naming for the most popular names has changed over time.

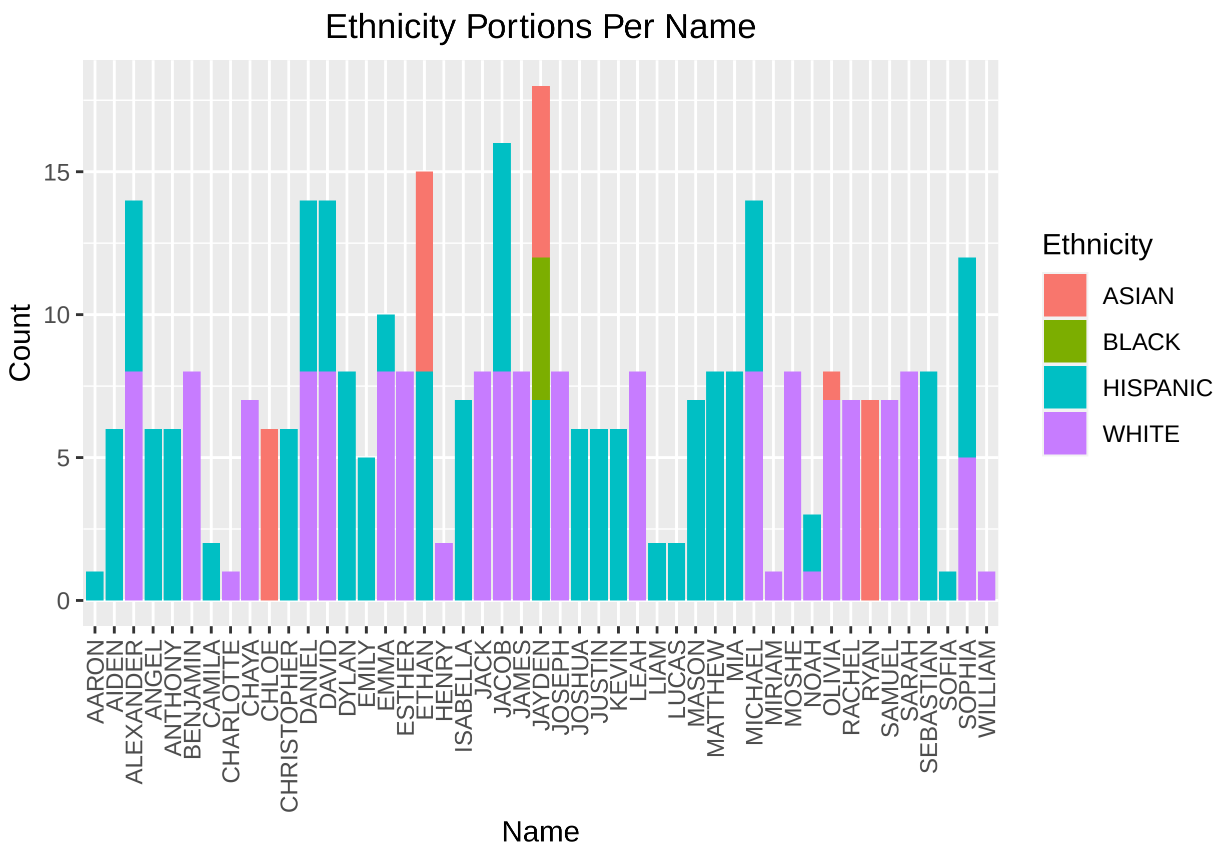
```
summary(names_data_by_count)
```

```
##       year         Gender           Ethnicity             name
##  Min.   :2011   Length:57582       Length:57582       Length:57582
##  1st Qu.:2012   Class :character   Class :character   Class :character
##  Median :2013   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2013
##  3rd Qu.:2014
##  Max.   :2019
##      Count             Rank              n
##  Min.   : 10.00   Min.   :  1.00   Min.   : 332
##  1st Qu.: 13.00   1st Qu.: 38.00   1st Qu.:2093
##  Median : 20.00   Median : 59.00   Median :3423
##  Mean   : 33.93   Mean   : 57.07   Mean   :2780
##  3rd Qu.: 36.00   3rd Qu.: 78.00   3rd Qu.:3914
##  Max.   :426.00   Max.   :102.00   Max.   :4334
```

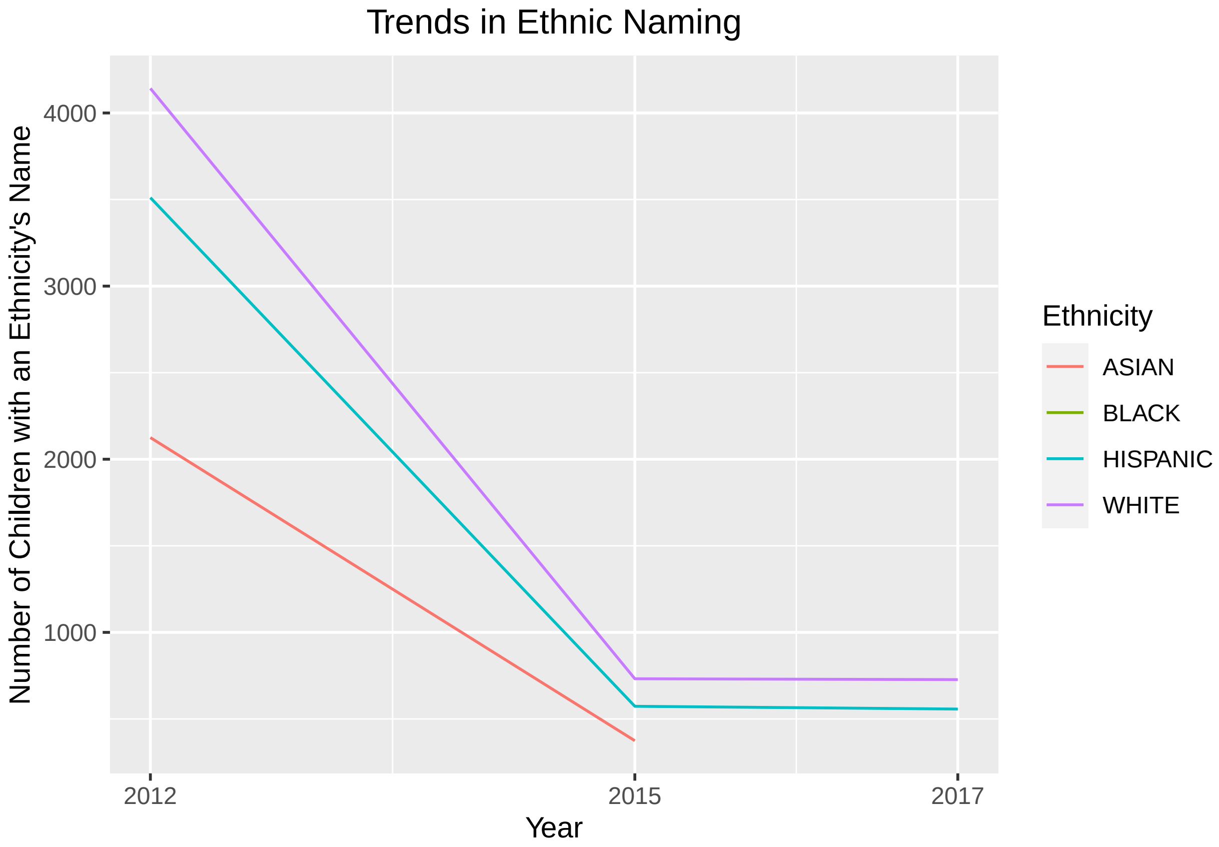
```
# graph to see change over time
ethnicities_blocked |>
  ungroup() |>
  group_by(Ethnicity) |>
  ggplot(aes(x = year,
             y = n,
             color = Ethnicity)) +
   geom_line() +
   labs(x = "Year",
        y = "Number of Names",
        title = "Trends in Ethnic Naming") +
   scale_x_continuous(breaks = labels) +
   theme(plot.title = element_text(hjust = 0.5))
```



```
# see proportions of names to ethnicity
names_by_year <- names_data |>
  add_count(year, name)

names_blocked <- names_by_year |>
  mutate(Ethnicity = case_when(Ethnicity %in% Asian ~ "ASIAN",
                               Ethnicity %in% Black ~ "BLACK",
                               Ethnicity %in% white ~ "WHITE",
                               TRUE ~ Ethnicity))

selected_years_names_blocked <- names_blocked |>
  filter(year %in% c(2012, 2015, 2017),
         Count > 170)

selected_years_names_blocked |>
  ggplot(aes(x = name, fill = Ethnicity)) +
    geom_bar() +
    labs(x = "Name",
         y = "Count",
         title = "Ethnicity Portions Per Name") +
    theme(plot.title = element_text(hjust = 0.5),
         axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



```r
ethnicities_blocked |>
  ungroup() |>
  group_by(Ethnicity) |>
  filter(year %in% c(2012, 2015, 2017),
         Count > 170) |>
  ggplot(aes(x = year,
             y = n,
             color = Ethnicity)) +
    geom_line() +
    labs(x = "Year",
         y = "Number of Children with an Ethnicity's Name",
         title = "Trends in Ethnic Naming") +
    scale_x_continuous(breaks = labels) +
    theme(plot.title = element_text(hjust = 0.5))
```



Results

Due to how ethnicities were determined and the general down trend of the most popular names, there are several theories that this data could support. One would be that as strict ethnicities become harder to determine, name ethnicity also becomes harder to determine.

Another interpretation would be that popular names are becoming more ubiquitous and less ethnicity-specific so popular names cannot be attributed to a single ethnicity which leads to the down trend (the names stay popular but no longer count as belonging to a specific ethnicity).

Contrarily, the graph could also be interpreted to mean that there is more name variation now (either in name spelling or completely different names) which leads to a decrease in popular names regardless of ethnic background.

The main conclusion that can be taken away from this project is that a more robust study would need to take place to clarify what has caused the trends seen here.

Notes

Ethnicity: determined by the mother's ethnicity

Source: New York's Department of Health and Mental Hygiene which maintains the birth registry

This data set is updated annually. It has data from 2012 to 2019.

Read more: *Streets of Gold: America's Untold Story of Immigrant Success* by Leah Boustan and Ran Abramitzky.