San Francisco Public Schools Lottery: An Analysis

Joseph Boyle

In keeping with my theme of playing around with data this summer, I wanted to examine an issue I had first heard about from my sister, who worked in San Francisco for several years. In the 1970s, the San Francisco Unified School District (SFUSD), like many other urban school districts, began busing its students in an effort to address segregation in its schools. This policy, and SFUSD's other efforts as integration, led many white students out of the district or into private schools, a phenomenon known as "white flight". Efforts at integration were not as effective as the district originally hoped. At the turn of the 21st century, SFUSD initiated a lottery system to assign students to schools based on a lottery, incorporating families' choices of schools in an algorithmic process.

This lottery has been the subject of intense scrutiny in the past years as parents have placed increasing importance on the schools their children attend, as well as school's features and amenities, racial composition, and underlying culture. Data from the lottery is public; I obtained it here, from San Francisco's local NPR station. This data covers the lottery assignment for kindergarteners in the 2017-2018 school year.

The algorithm gives tie-breaker preference to students living in zones in the city with lower test scores (CTIP-1 zones), students with siblings already in a chosen school, and students living in the given attendance zone for a chosen school. Beyond that, the algorithm reportedly evaluates students' choices by school, filling up a school with empty spots after these tiebreakers randomly with students who had chosen it, and proceeding until all schools are filled. Consequently, the algorithm chooses random pairs of students, evaluates whether both students would be better off (with respect to their families' choices) if they switched schools, and switches them if this is the case. At this point, assignments are mailed to families. There are more rounds in the algorithm available, but the chance of improving a student's assignment becomes markedly lower after the first round; most students either go to their assigned school or leave the district and attend private school.

Some articles claim that white students receive disproportionately positive outcomes in the lottery process. Despite the fact that confidentiality concerns in the public data prevent me from simulating the algorithm's process as closely as possible, I set out to evaluate these claims, and model, visualize, and analyze the process of the SFUSD kindergarten lottery!

Getting Started

Reading in the data and loading packages

```
library(tidyverse)
library(ggmap)
library(MASS)
library(magrittr)
library(knitr)

sf <- read.csv("sf_kindergarten.csv")
sf_codes <- read.csv("sf_codes.csv")
sf_income <- read.csv("sf_income.csv")
```

The sf_income dataset includes median household income figures by zip code from the 2016 American Community Survey. I considered using this as a covariate in modeling results from the lottery process, specifically the ordinal choice ranking for a student's school assignment. Since one observation in the dataset originated in Sacramento County, I used their county median household income.

Cleaning

I did a bit of cleaning, making the column names a bit nicer and editing the column types. At least one parent (I'm using this as a catch-all term to signal those who made the school choices) ranked all 92 possibilities for kindergartens in their child's form, so the ordinal choice variables range from c1 to c92, with NA values if a parent did not rank that many choice:

```
colnames(sf) <- c("id", paste("c", 1:92, sep = ""),
                  "assignment", "enrolled",
                  "race", "ctip1", "zip")

sf$race <- as.factor(sf$race)
sf$zip <- as.factor(sf$zip)
```

Counting

Next, I counted the number of choices that each parent listed, creating a variable for this:

```
# For each student in the dataset,
for (i in 1:nrow(sf)){

  # Initialize an empty vector for the student
  count = numeric()

  # For each column, from c1 to c92,
  for (j in 2:93){

    # If there is a non-NA value in the column,
    if (!is.na(sf[i, j])){

      # Add one to the number of ranked choices...
      count <- c(count, 1)}
  }

  # And, upon finishing, state the count variable's value as
  # the number of choices in the student's ranking form
  sf$count[i] <- sum(count)

}
```

Checking the distribution of number of ranked choices, there is a relatively small number of parents having listed many schools. To slightly simplify matters, I will only consider the lower 95% of the number of ranked schools, speeding things up while realistically losing a very small amount of information. Afterwards, I re-run the choice-counting loop with the trimmed data:

```
quant <- quantile(sf$count, .95) %>%
         as.numeric()
```

(Although the value of quant was not outputting for me, it is 40!)

```
sf <- sf %>% dplyr::select(-(c41:c92))
```

```
for (i in 1:nrow(sf)){

  count = numeric()

  for (j in 2:41){

    if (!is.na(sf[i, j])){

      count <- c(count, 1)}
  }

  sf$count[i] <- sum(count)

}
```

Ordinal Response Variable

After just a bit more cleaning, I joined the income data to the school choice data and created an ordinal response variable, indicating the assigned school's rank in the parents' ranking list.

```
sf_income$zip <- as.factor(sf_income$zip)

sf <- left_join(sf, sf_income, by = "zip")
```

```
# For each student in the dataset,
for (i in 1:nrow(sf)){

  # Note which school the student was assigned to
  assignment <- sf$assignment[i]

  # For each choice ranking,
  for (j in 2:41){

        # If the ranking matches the assignment,
        if (sf[i, j] == assignment){

          # The "ordinal" variable is set to the ordinal choice rank
          sf$ordinal[i] <- j - 1}
  }
}
```

Joining datasets

I did a lot of hard-coding school addresses from the SFUSD website, the details of which I'll spare, but my next step was to join the school data, now complete with addresses, to the school choice data. Now, we have geographic information in the school choice data for where the students were assigned:

```
sf <- left_join(sf, sf_codes, by = c("assignment" = "code"))

sf <- sf %>% dplyr::select(-name, -school_zip, -address)

colnames(sf)[48] <- c("zip_income")
colnames(sf)[55] <- c("assigned_address")
```

Geocoding, and joining again

To get a spatial representation - with GPS coordinates - of where the schools were in the district, I used the geocode() function from the ggmap package. For me, geocode() ended up being buggy, denying requests and the like, so I had to go through a few iterations, as well as hard-code the GPS coordinates for a few schools that the function messed up.

```
# Geocode assigned schools and also name them

unique_addresses <- unique(sf$assigned_address)

unique_geocodes <- geocode(unique_addresses)

unique_addresses_and_geocodes <- cbind(unique_addresses, unique_geocodes)


# Update school choice dataframe to reflect GPS coordinates of school student is assigned to
sf <- left_join(sf, unique_addresses_and_geocodes,
                by = c("assigned_address" = "unique_addresses"))

sf <- left_join(sf, sf_codes,
                by = c("assignment" = "code")) %>%
      dplyr::select(-school_zip, -address, -full_address)
```

Mapping!

To visualize where these students are being assigned, I wrote a function that mapped the path of each student in a given zip code. Since zip-code-level was as granular as the dataset would allow, the plot for each zip code originates all students from the geocode() coordinates for the zip code; lines from there are mapped to the coordinates of the school to which they were assigned. To reduce overplotting, I set the alpha argument of geom_segment to be darker if a greater proportion of students from that zip code were assigned to a given school, and lighter if fewer were.

```
# Set themes and produce base map of SFUSD
theme_set(theme_bw(15))

map <- get_map(location = c(lon = -122.45, lat = 37.75),
               source = "stamen", maptype = "toner-lite",
               zoom = 12, color = "bw")

# Set location of zip code label on map
text_lon <- -122.5; text_lat <- 37.8

# Function, taking input for a given zip code
```

```r
# function, taking input for a given zip code

mapper <- function(the_zip){

  # Incorporate and geocode the zip code
  zip <- as.character(the_zip)

  zip_gps <- geocode(zip)

  zip_text <- data.frame(
                  name = zip,
                  text_lon = text_lon,
                  text_lat = text_lat
                     )

  # Set the longitude and latitude for the zip code
  # (Place in data frame due to ggplot requirement)
  d <- data.frame(
                  lon = as.numeric(zip_gps[1]),
                  lat = as.numeric(zip_gps[2])
                 )

  # From where will the segments originate?
  start_lon <- d[1,1]
  start_lat <- d[1,2]

  # Set up data frame for students in a given zip code.
  # Group by their assigned school.
  zip_assignments <- sf %>%
                     filter(zip == the_zip) %>%
                     group_by(assignment)

  # How many of them are assigned to a given school?
  zip_assignments <- zip_assignments %>%
                     mutate(n = n()/nrow(zip_assignments)) %>%
                     distinct(assignment, .keep_all = T) %>%
                     select(assignment, name, n, lon, lat) %>%
                     mutate(start_lon = d[1,1]) %>%
                     mutate(start_lat = d[1,2])

  # ggplot call
  map <- ggmap(map) +

      # Add a segment to the map for each school that a student is assigned to
      # It will end at the coordinates of the school, and its shading will be
      # positively associated witht the proportion of students going to that school
      geom_segment(data = zip_assignments,
                   aes(x = start_lon, y = start_lat,
                       xend = lon, yend = lat,
                       size = 2.5,
                       alpha = n/max(n)),
                   color = "gold",
                   lineend = "round") +

      scale_size(range = c(0.5, 2.5)) +

      # Add a label for the originating zip code
      geom_text(data = zip_text,
                mapping = aes(label = name,
                              x = text_lon,
                              y = text_lat),
                size = 9) +

      theme(axis.line = element_blank(),
            axis.text = element_blank(),
            axis.ticks = element_blank(),
            legend.position = "none") +
      xlab("") +
      ylab("")

    # Save plot to my computer
    ggsave(map, device = "pdf", width = 14, height = 7, units = "in",
           file = paste0("map_", zip, ".pdf"))
```

```
    # Incorporate a dummy while loop so that geocode() doesn't bark at me
    # for querying too many conversions in not enough time
    date_time<-Sys.time()

    while((as.numeric(Sys.time()) - as.numeric(date_time)) < 30) {}

}
```

Now that this function exists, we can apply it to each zip code in the dataset:

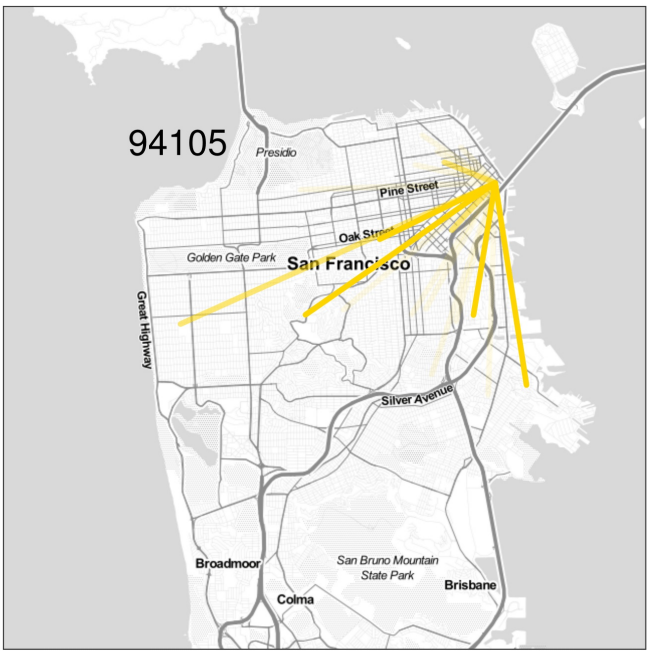
```
for (i in 1:length(unique(sf$zip))){
  mapper(unique(sf$zip)[i])
}
```

Maps

Below are the plots produced for each of the zip codes in the dataset. Now, we can see what parts of the district students from a given zip code are assigned to, again with darker lines corresponding to more students. (I removed plots from zip codes 94030, 94066, 94080, 94401, 94538, 94801, and 94806 because of the extremely low number of students applying from those zip codes, which produced essentially an empty plot.)


94103
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

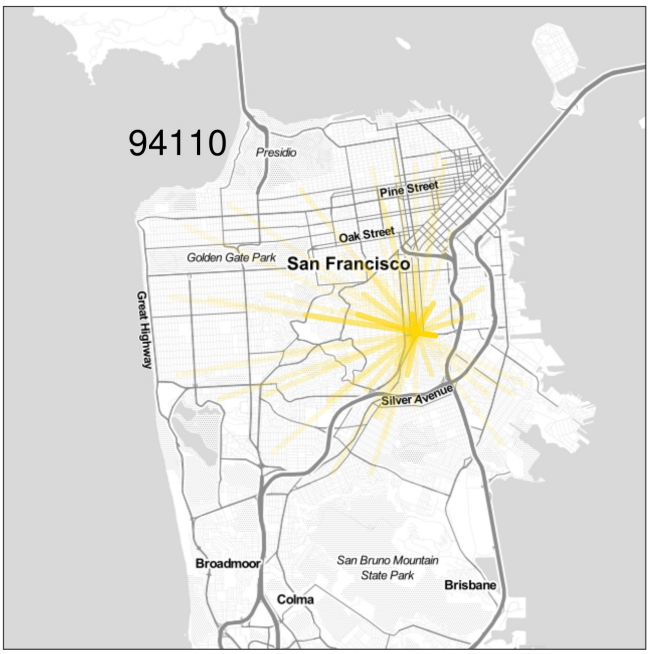

94104
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94105
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

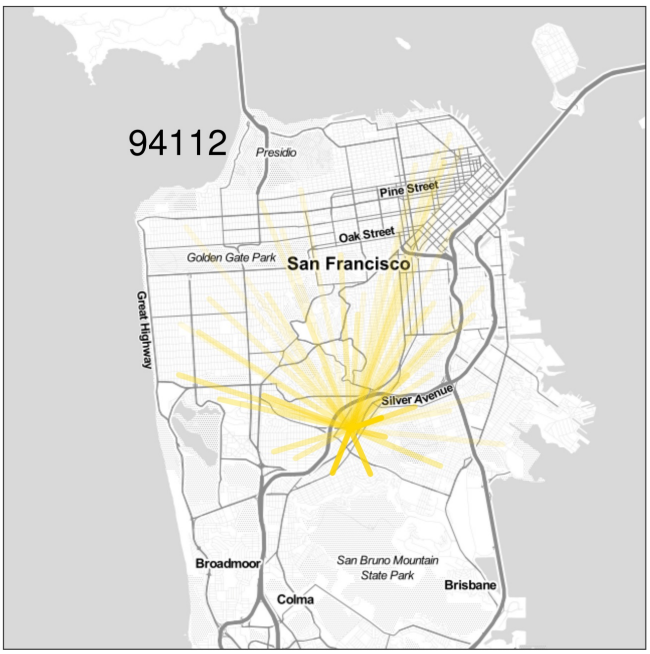

94107
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

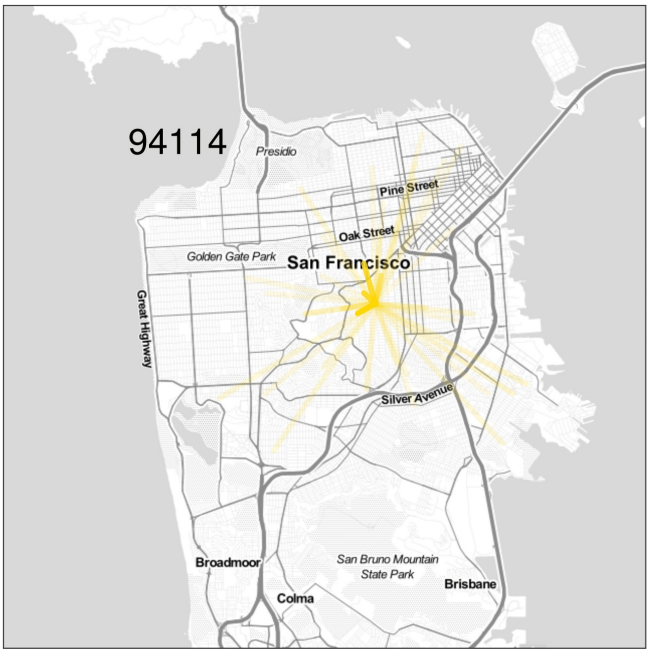

94108
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

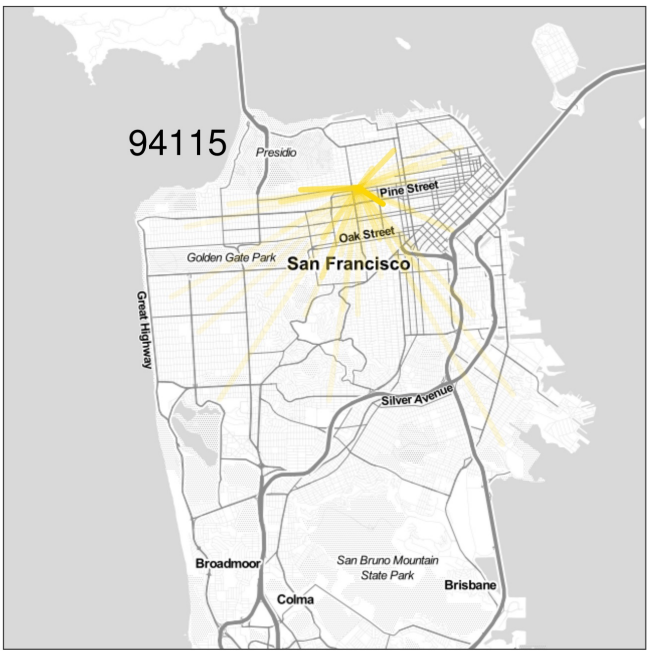

94109
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

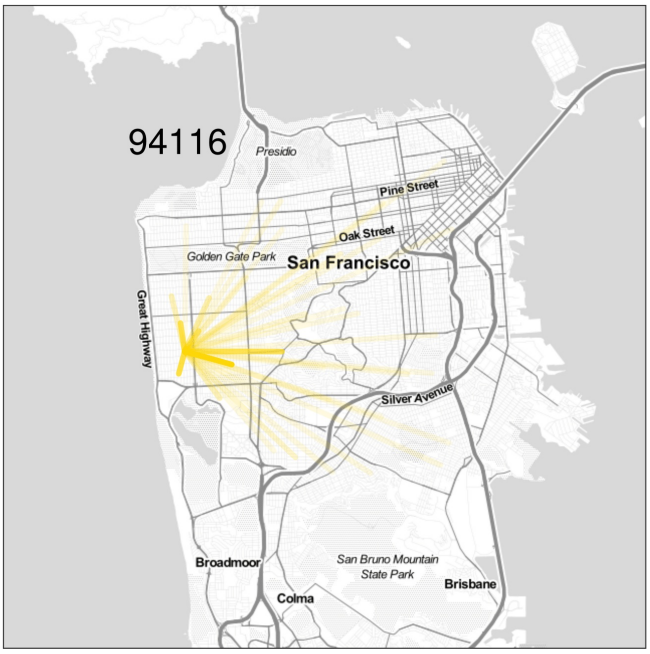

94110
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

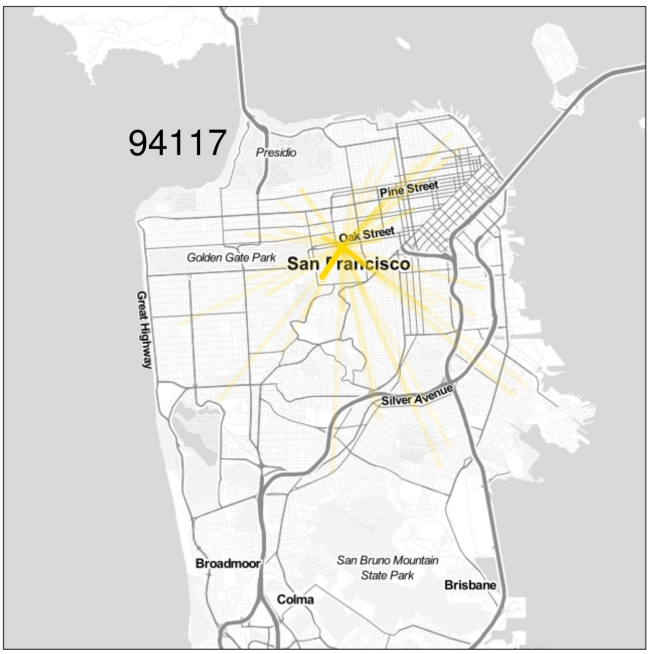

94111
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

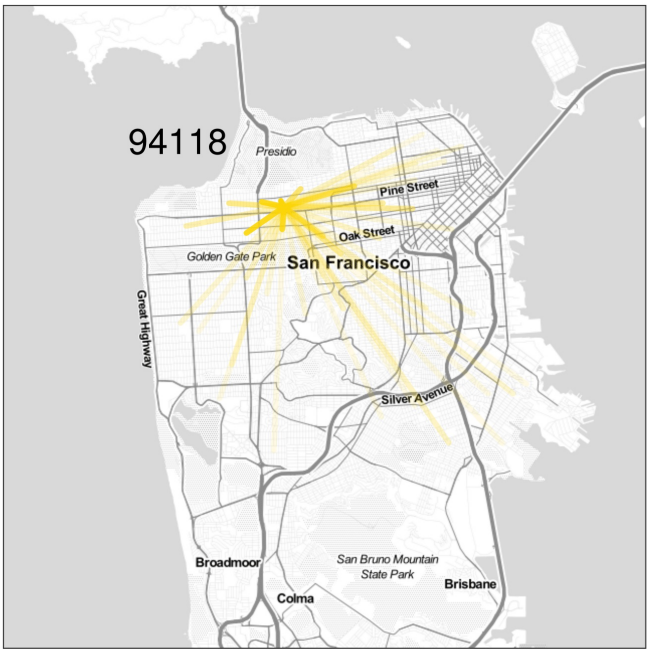

94112
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94114
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

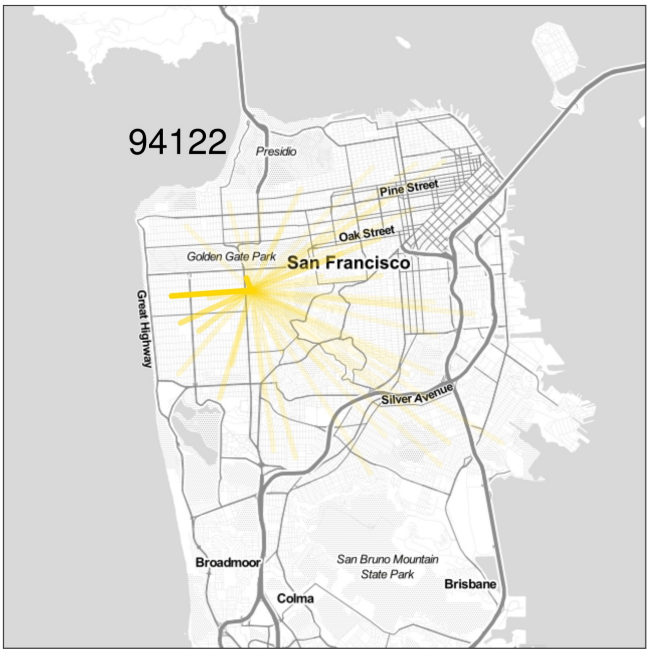

94115
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94116
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

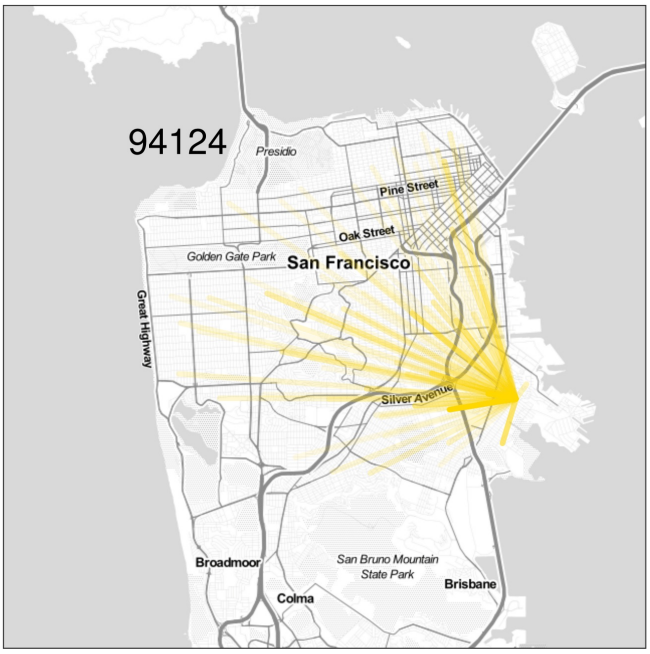

94117
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

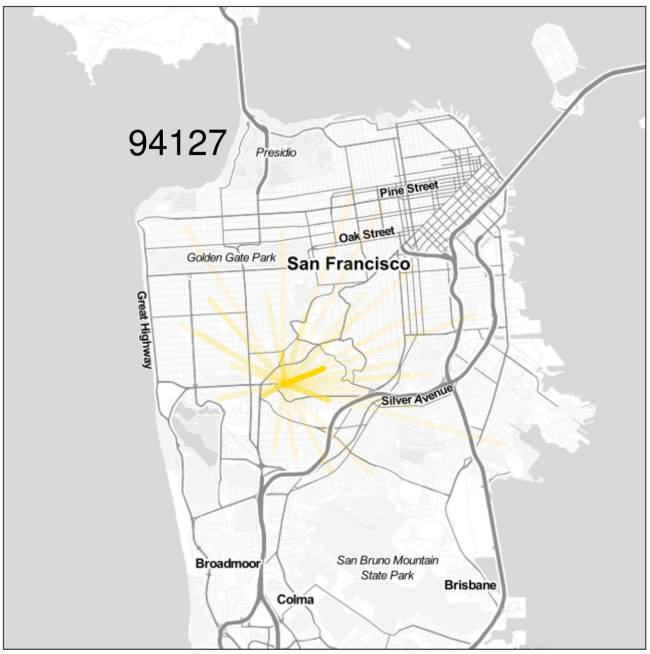

94118
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94121
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94122
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94123
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

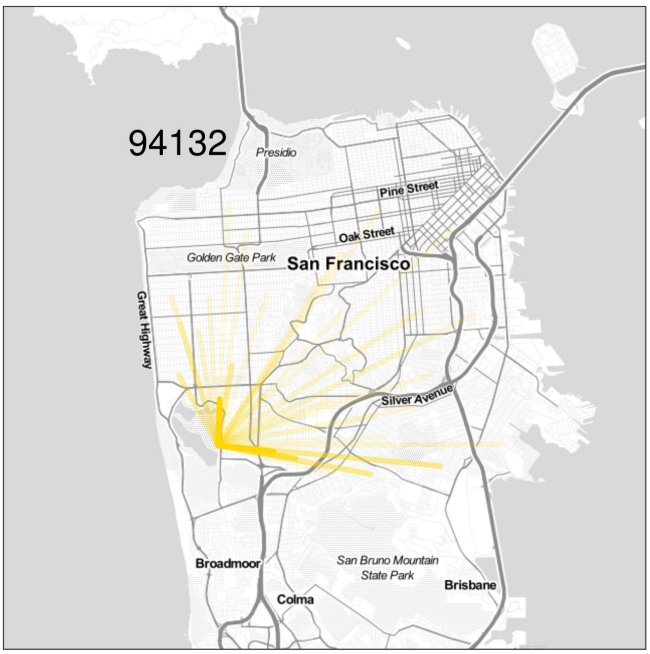

94124
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

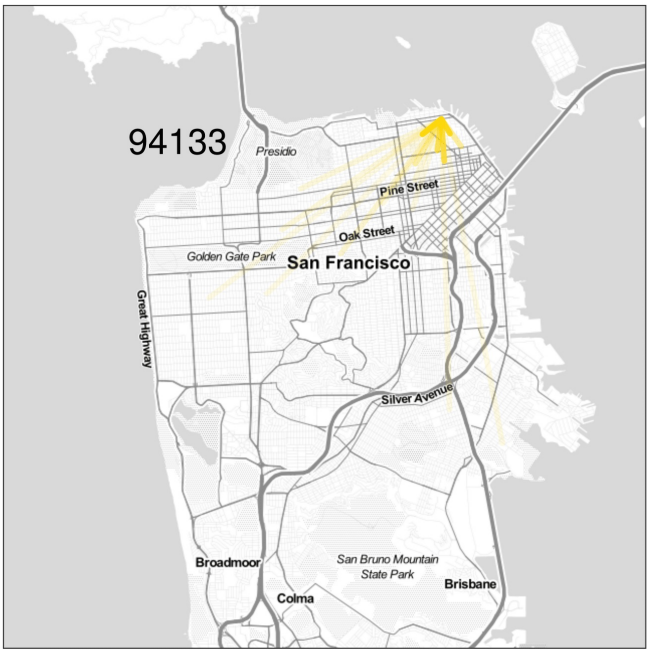

94127
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma

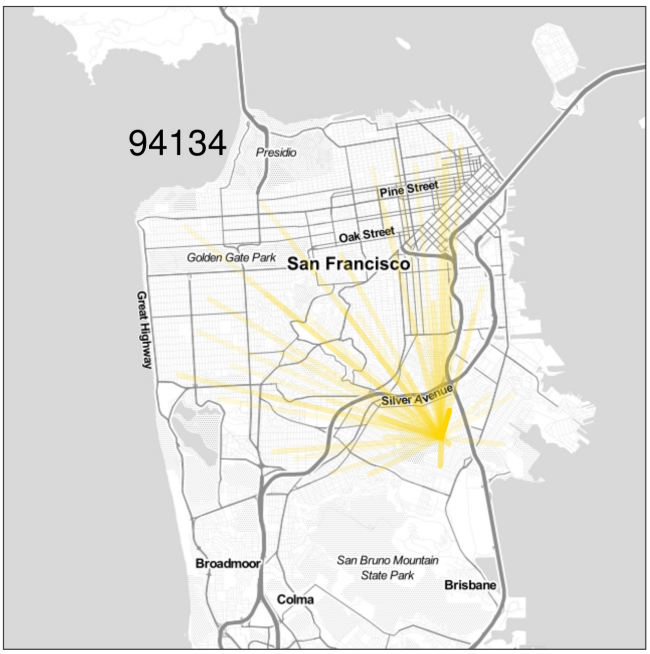

94129
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94130
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94131
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94132
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma


94133
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
Broadmoor
San Bruno Mountain
State Park
Brisbane
Colma



Some things that stood out at me from these plots: the assignments of

- many students in **94103**, a zip code with a relatively **high** proportion of Black and Asian residents, to schools **far** away from the zip code
- many students in **94105**, a zip code with a relatively **high** proportion of White and Asian residents, a relatively **low** proportion of Black and Hispanic residents, and a median household income of $164,000, the **fifth-highest** of the 183 zip codes in the metropolitan area, to schools **far** away from the zip code
- students in **94110**, a densely populated zip code incorporating the Mission District, seemingly **uniformly** around the city; looking at the ordinal ranking values, however, 60 percent of students were assigned to their first choice, and 76 percent were assigned to one of their first two choices
- students in **94115**, a zip code adjacent to the Presidio highly populated with White residents, to schools largely very **close** to the zip code
- students in **94121**, a high-income zip code encompassing Richmond and Outer Richmond, to schools largely very **close** to the zip code; additionally, 67 percent of students were assigned to their first choice, and 79 percent were assigned to one of their first two choices
- many students in **94124**, a low-income zip code in the south-eastern corner of the district, largely consisting of elderly Black and Asian residents, to schools **far** from the zip code

There's definitely a lot to see and think about here, but I'd like to do a few other things before trying to draw conclusions.

School Desirability

Given how the school lottery algorithm is reported to work, it makes sense that the “desirability” of a given school plays some role in the outcomes of those parents who list it on their form. Simply put, if everyone wants their child to go to a school, some will be left disappointed; it can only enroll so many. I sought to quantify this desirability for my modeling of the process.

I figured that there were two main aspects I should consider in quantifying this - “popularity”, which I decided to measure by evaluating whether it was on parents’ top-10 choices, and desirability, as measured by its median ranking across all choice forms in the dataset. The code for these measures is below.

```
# Create a "popularity" metric for schools

sf_codes$popularity <- numeric(nrow(sf_codes))

# For each school:
for (i in 1:nrow(sf_codes)){

  # Consider the school in question; initialize an empty vector
  school_code <- sf_codes$code[i]
  choice_counter <- numeric()

  for (j in 1:nrow(sf)){

    # If a parent lists the school in their top-10 choices,
    for (k in 2:11){
      if (sf[j, k] == school_code){

        # Make note of this and
        choice_counter <- c(choice_counter, 1)}
    }
  }

  # Tally the number of these occurences
  sf_codes$popularity[i] <- sum(choice_counter)
}
```

```
# Create a "desirability" metric for schools

sf_codes$school_desire <- numeric(nrow(sf_codes))

# For each school:
for (i in 1:nrow(sf_codes)){

  # Consider the school in question; initialize an empty vector
  school_code <- sf_codes$code[i]
  school_vec <- numeric()

  for (j in 1:nrow(sf)){

    # If a parent ranks the school,
    for (k in 2:41){
      if (sf[j, k] == school_code){

        # Make note of the ordinal ranking and
        school_vec <- c(school_vec, k-1)}
    }
  }

  # Find the median value of these rankings
  sf_codes$school_desire[i] <- median(school_vec)
}

# And replace any NA values with the most unpopular value of this metric
max_unpopularity <- max(sf_codes$school_desire[!is.na(sf_codes$school_desire)])

sf_codes$school_desire[is.na(sf_codes$school_desire)] <- max_unpopularity
```

For ease of comparison, I converted both of these measures to percentiles and computed their average to be my desirability metric for the schools. This took into account how many parents were choosing a school, as well as how highly they were coveting it.

```
# Convert these metrics to percentiles

sf_codes$popularity_percentile <- (sf_codes$popularity - min(sf_codes$popularity)) /
  (max(sf_codes$popularity) - min(sf_codes$popularity))

sf_codes$desirability_percentile <- (max(sf_codes$school_desire) - sf_codes$school_desire) /   (max(sf_codes
$school_desire) - min(sf_codes$school_desire))

# Average percentiles for composite metric

sf_codes$desire <- 0.5 * (sf_codes$popularity_percentile + sf_codes$desirability_percentile)
```

And I added this to the dataset.

```
# Append metric to dataframe

desire <- sf_codes %>% dplyr::select(code, desire)

sf <- left_join(sf, desire, by = c("c1" = "code"))
```

Modeling

The context in which I had been thinking about this process has been ordinal - what is the choice of ranking of the school that the lottery assigns a student to? Therefore, linear regression is inappropriate - the range of values of {1, 2, 3, …}, certainly not continuous and possibly not with equal spacings in real-world value between numerical values. I modeled the process with ordinal regression: specifically, the polr() function from the MASS library. The function is so named because it fits a proportional-odds logistic regression model.

Guided by the vignettes here, I produced an ordinal regression modeling the ordinal choice rank to which a student was assigned. I went about this process via forward selection of variables, with variable p-value < 0.05 and minimizing AIC as the advancement criteria. The code for my final model, and its summary, follow.

```
# Produce ordinal regression model and store coefficients
model <- polr(ordinal ~ ctip1 + desire + hispanic + black,
              data = sf, Hess = T)

ctable <- coef(summary(model))

# Find p-values for each model coefficient, based on t-value
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2 %>% round(2)

# Find odds ratios for  coefficients
odds_ratios <- exp(ctable[, "Value"]) %>% round(2)

ctable <- cbind(ctable, "p value" = p, odds_ratios) %>% head(4)
```

Again, I was not able to capture every “tiebreaker” the algorithm takes into account - if the student in question has a sibling already going to a chosen school, and/or lives in the attendance area for that school. However, with the data I did have, the following model tells a decent amount of the story.

```
##               Value Std. Error    t value      p value odds_ratios
## ctip1Y    1.5322258 0.10373229  14.770963 2.254690e-49        4.63
## desire   -3.3811715 0.19032536 -17.765218 1.314290e-70        0.03
## hispanic  0.1640565 0.07246584   2.263915 2.357934e-02        1.18
## black     0.4291409 0.20550346   2.088242 3.677605e-02        1.54
```

What these coefficients, with their corresponding odds ratios, tell us are the marginal effects of having these characteristics on the school assignment’s ordinal choice ranking, conditioned on the other variables in the model. So, all else equal:

- living in a CTIP-1 zone makes a student 4.63 times more likely to be assigned to their n-th best choice, as opposed to their (n+1)-th best choice,
- listing a first-choice school one unit higher on the aforementioned desire metric makes a student 0.03 times as likely to make this jump (it is worth noting, however, that the desire metric takes values in [0, 1], so a more relevant interpretation of this is that listing a first-choice school that is moderately more desirable than another school makes a student 0.18 times as likely, or listing one that is somewhat more desirable makes a student 0.43 times as likely,
- being of Hispanic ethnicity makes a student 1.18 times more likely to make this jump, and
- being Black makes a student 1.54 times more likely to make this jump.

Of course, the story would be much more complete if I had access to sibling and living-in-attendance-area data, but there is no reason that I should have access to that!

Evaluating the Washington Post's claim

The Washington Post article I linked to in the introduction claims that "white kids are winning San Francisco's school lottery, and the data proves it" (3/27/2015). Their main argument here is that the lottery "had the effect of concentrating white students in the best elementary schools", something they apparently synonymize with higher standardized test scores. The author provides a scatterplot of schools' percentage enrollment of white students plotted against their Academic Performance Index, with a displayed R^2 just over 20 percent. This appears to be the basis of their claims.

My first thought upon seeing this was that the claim was not necessarily matching the evidence. To me, this seemed to be saying that the more white children a school enrolled, the higher the school's test scores were - and nothing else; their analysis did not take into account any other school-related factors that parents value.

To this end, I wanted to answer their question in what I consider a more accurate way - are kids of any race or ethnicity "winning" the lottery? To do this, I looked at the five elementary schools scoring highest on the desirability index: Clarendon, West Portal, Grattan, Rooftop, and Lilienthal.

For each of these schools, I compared the actual enrollment by race to how strongly the school was prioritized by race: was it a top-five choice in the parents' ranking? The "difference" denotes the percentage students of a given race were assigned to a school more than would have been expected by their parents' prioritizing. (Thus, negative values denote percentages less likely than would have been expected.) The code for this process is below for Clarendon; I repeated it for the four other schools afterward.

```
# Actual Enrollment by race

clarendon <- sf %>%
             filter(enrolled == 478) %>%
             dplyr::select(race)

# Create table with proportions
clarendon <- cbind(table(clarendon$race)/nrow(clarendon))

# Theoretical by race, based on choices

clarendon_choice <- sf %>%
                    filter(c1 == 478 | c2 == 478 | c3 == 478 | c4 == 478 | c5 == 478) %>%
                    dplyr::select(race)

# Create table with proportions
clarendon_choice <- cbind(table(clarendon_choice)/nrow(clarendon_choice))

# Combine

clarendon_table <- cbind(clarendon, clarendon_choice)
colnames(clarendon_table) <- c("enrolled", "prioritized")

clarendon_table[, 1] <- round(clarendon_table[, 1], 2)
clarendon_table[, 2] <- round(clarendon_table[, 2], 2)

clarendon_table[, 1] <- as.numeric(clarendon_table[, 1])
clarendon_table[, 2] <- as.numeric(clarendon_table[, 2])

clarendon_difference <- clarendon_table[, 1] - clarendon_table[, 2]
clarendon_table <- cbind(clarendon_table, clarendon_difference)
colnames(clarendon_table)[3] <- c("difference")
```

This produces the following set of tables by race for the most highly desired elementary schools in the district:

```
clarendon_table
```

```
##                 enrolled prioritized difference
## asian               0.24        0.18       0.06
## black               0.01        0.01       0.00
## hispanic            0.11        0.12      -0.01
## native american     0.01        0.00       0.01
## other               0.27        0.29      -0.02
## white               0.36        0.39      -0.03
```

```
wp_table
```

```
##          enrolled prioritized difference
## asian        0.39        0.43      -0.04
## black        0.01        0.02      -0.01
## hispanic     0.08        0.09      -0.01
## other        0.28        0.27       0.01
## white        0.23        0.19       0.04
```

```
grattan_table
```

```
##          enrolled prioritized difference
## asian        0.24        0.11       0.13
## black        0.01        0.03      -0.02
## hispanic     0.11        0.14      -0.03
## other        0.27        0.23       0.04
## white        0.36        0.50      -0.14
```

```
rooftop_table
```

```
##          enrolled prioritized difference
## asian        0.14        0.08       0.06
## black        0.04        0.04       0.00
## hispanic     0.26        0.17       0.09
## other        0.17        0.23      -0.06
## white        0.38        0.47      -0.09
```

```
lili_table
```

```
##          enrolled prioritized difference
## asian        0.20        0.15       0.05
## black        0.02        0.03      -0.01
## hispanic     0.11        0.11       0.00
## other        0.30        0.27       0.03
## white        0.36        0.45      -0.09
```

To me, this says the opposite of what the Washington Post article claims. If “winning” the lottery means assignment to the most highly-desired public schools, then white students do not appear to be “winning”" in this sense. If anything, Asian students appear to be “winning” the lottery! The picture is more complicated than it may initially seem.

Conclusions

This was absolutely an interesting dataset to examine. The subject is vital to so many Bay Area parents and guardians, who, according to many of the articles I had read, seem to place a significant amount of importance on the schools (even elementary schools, as here!) to which their children are assigned.

I knew from the outset that, in the absence of all the information that could best simulate the algorithm’s performance - sibling and attendance zone data - the granularity with which I could depict the story was limited. So, I decided to take a few different paths: mapping the weight of students’ assignments from a given zip code, modeling the ordinal ranking choice to which a student was assigned, and comparing the enrollment data for the “most popular” schools with the families who had prioritized attending them.

One thing that stood out to me was that the CTIP-1 tiebreaker preference seems to hold water very well. Students from census tracts with low average test scores are assigned to schools that their parents feel are best for them; if this means sacrificing convenience for resources and school culture, these parents are giving a thumbs-up, and the algorithm is obliging them. This is a good thing. Another thing I noticed is that, in areas with extremely high incomes, and presumably high-quality schools, students are being assigned to these neighborhood schools, preserving these pockets of wealth and educational might. Finally, I think that some parents would be better off if they were more realistic about their child’s chances of “winning” assignment to highly-sought-after elementary schools, if there is no reason that they would be assigned to them. If they are not in the attendance zone, not in a CTIP-1 zone, and/or don’t already have a sibling attending the school, the model and popular school comparisons suggest that there is little use in placing ranking hopes and dreams on Grattan, West Portal, or the like.

There is so much more that can be done with this data. The first thing that comes to mind for me is investigating clustering in school preferences by certain characteristics, such as race.