NYC Food Environment and Health

Gaeun Kim

Motivation

In 2017, a report from Centers for Disease Control and Prevention revealed that America's obesity rate has reached a record high. In contrast to the popular belief, New Yorkers are not so safe from the obesity epidemic, as more than half of adult New Yorkers are either overweight or obese. Studies show that the rise in the obesity epidemic is partly due to disparities in food environment; it is harder for some to eat healthier because their options are limited.

This project intends to look deeper into the relationship between food environment in NYC and obesity rate along with diabetes rate and stroke hospitalization rate.

Initial questions

1. Can the percentage of nation-wide chain and fast-food restaurants in different neighborhood of NYC be related to their obesity, diabetes and stroke hospitalization rate?

2. Can the percentage of fast-food restaurant (defined by cuisine description in the restaurant inspection dataset) in different neighborhood of NYC be related to their obesity, diabetes and stroke hospitalization rate?

3. Can the percentage of health inspection grade A restaurant in different neighborhood of NYC be related to their obesity, diabetes and stroke hospitalization rate?
4. Can the composite restaurant health score of different neighborhood of NYC be related to their obesity, diabetes and stroke hospitalization rate?

For the first two questions, we stick to it. For the third one, After looking into the backgroud of the restaurant inspection data, we found out that the percentage of health inspection grade A restaurant is intuitively not related to the three chronic disease health outcomes, so we later consider it as a confounder. We also deleted the fourth problem, since composite restaurant health score is too subjective and difficult to assess. Moreover, we fitted models on the borough level first due to the difficulties we encountered in scaping zipcode-neighborhood data. However, after some struggle, we manage to get the neighborhood information and analyze the data accordingly.

Data and Methods

Data Source and Collection

1. NYC restaurant inspection:

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

This dataset contains the data for restaurant inspection in NYC from August 1, 2014 to June 9, 2019. Every row is a restaurant inspection record that includes the name of the restaurant, zipcode, cuisine type description, inspection grade (A as the best grade) and so on.

Variable used in this datasets

DBA: Name of the restaurant
BORO: name of the boro
ZIPCODE: the zipcode of the restaurant
CUISINE DESCRIPTION: the kind of food that the restaurant is providing
GRADE: Inspection grade for the specific inspection.

2. 2015 Community Health Profiles Open Data:

https://www1.nyc.gov/site/doh/data/data-publications/profiles.page

This dataset contains NYC every neighborhoods' demographic (percentage of white race, poverty percentage), health (age-adjusted percent of adult exercised in the last 30 days, age-adjusted percent of adults as a smoker) and our main outcome (Age-adjusted percent of adults that is obese (BMI of 30 or gthe reater), Age-adjusted percent of adults, Age-adjusted rate of hospitalizations due to stroke (cerebrovascular disease) per 100,000 adults)

Variable used in this datasets

Name: the name for the neigborhood.
Racewhite_Rate: the percentage for white race
Poverty: Percent of individuals living below the federal poverty threshold
Smoking: age-adjusted percent of adults as a smoker Exercise: Age-adjusted percent of adults that reported getting any exercise in the last 30 days
Obesity: Age-adjusted percent of adults that is obese (BMI of 30 or greater) based on self-reported height and weight
Diabetes: Age-adjusted percent of adults that had ever been told by a healthcare professional that they have diabetes
Stroke_Hosp: Age-adjusted rate of hospitalizations due to stroke (cerebrovascular disease) per 100,000 adults

3. Web scraping for what zipcodes each neighborhood contains

Because Community Health Profiles Open Data has only neighborhood level information, so we have to aggregate neighborhood level information from the restaurant inspeciton data. However, the restaurant inspection data is using zipcodes for every restaurant. Therefore, we have to add the neighborhood name to every restaurant, so that we can group by neighborhood then calculate the percentage for every neighborhood.

First, we scraped the table data from https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm. However the neighborhood name and combinations are different from the health profile data we downloaded. We could not merge the two datasets. Then we tried to look for the raw classification for the neigborhoods in health data from New York University's Furman Center for Real Estate and Urban Policy and the NYC Department of City Planning. However, still no luck becuase the neighborhood is not divided by zipcode areas. Finally we tried the hardest way: searching the name of the neighborhood on Google and finding the matching zipcodes. Although it is not precise because neighborhoods are not divided according to the area of the zipcode (different neighborhoods sometimes share the same zipcodes), it works out well.

To deal with this issue and get an unbiased result, we randomly assign the zipcode to only one neighborhood if two or more neighborhoods are in the same zipcode area.

Data Import and Cleaning

First, we will download restaurant inspection data from NYC open data.

```
get_all_inspections = function(url) {

  all_inspections = vector("list", length = 0)

  loop_index = 1
```

```r
  chunk_size = 50000
  DO_NEXT = TRUE

  while (DO_NEXT) {
    message("Getting data, page ", loop_index)

    all_inspections[[loop_index]] =
      GET(url,
          query = list(`$order` = "zipcode",
                       `$limit` = chunk_size,
                       `$offset` = as.integer((loop_index - 1) * chunk_size)
                       )
          ) %>%
      content("text") %>%
      fromJSON() %>%
      as_tibble()

    DO_NEXT = dim(all_inspections[[loop_index]])[1] == chunk_size
    loop_index = loop_index + 1
  }

  all_inspections

}

url = "https://data.cityofnewyork.us/resource/9w7m-hzhe.json"

rest_inspection = get_all_inspections(url) %>%
  bind_rows()
```

```r
# changing to date
rest_inspection = rest_inspection %>%
  mutate(inspection_date = rest_inspection$inspection_date %>%
  strtrim(., nchar(.)-13) %>%
  as.Date() )
rest_inspection$inspection_date %>%
  max()
```

```
## [1] "2019-06-08"
```

Then, we will download the health and demographic data CSV into our local folder and clean it.

```r
download.file("https://www1.nyc.gov/assets/doh/downloads/excel/episrv/2015_CHP_PUD.xlsx", mode="wb", de
health <- read_excel("health.xlsx", sheet = "CHP_all_data") %>%
  select(Name, Racewhite_Rate, Poverty, Unemployment,
         Smoking, Exercise,
         Obesity, Diabetes, Stroke_Hosp, Airquality_rate) %>%
  clean_names()
```

Lastly, we will match the neighborhood names with restaurant zipcodes.

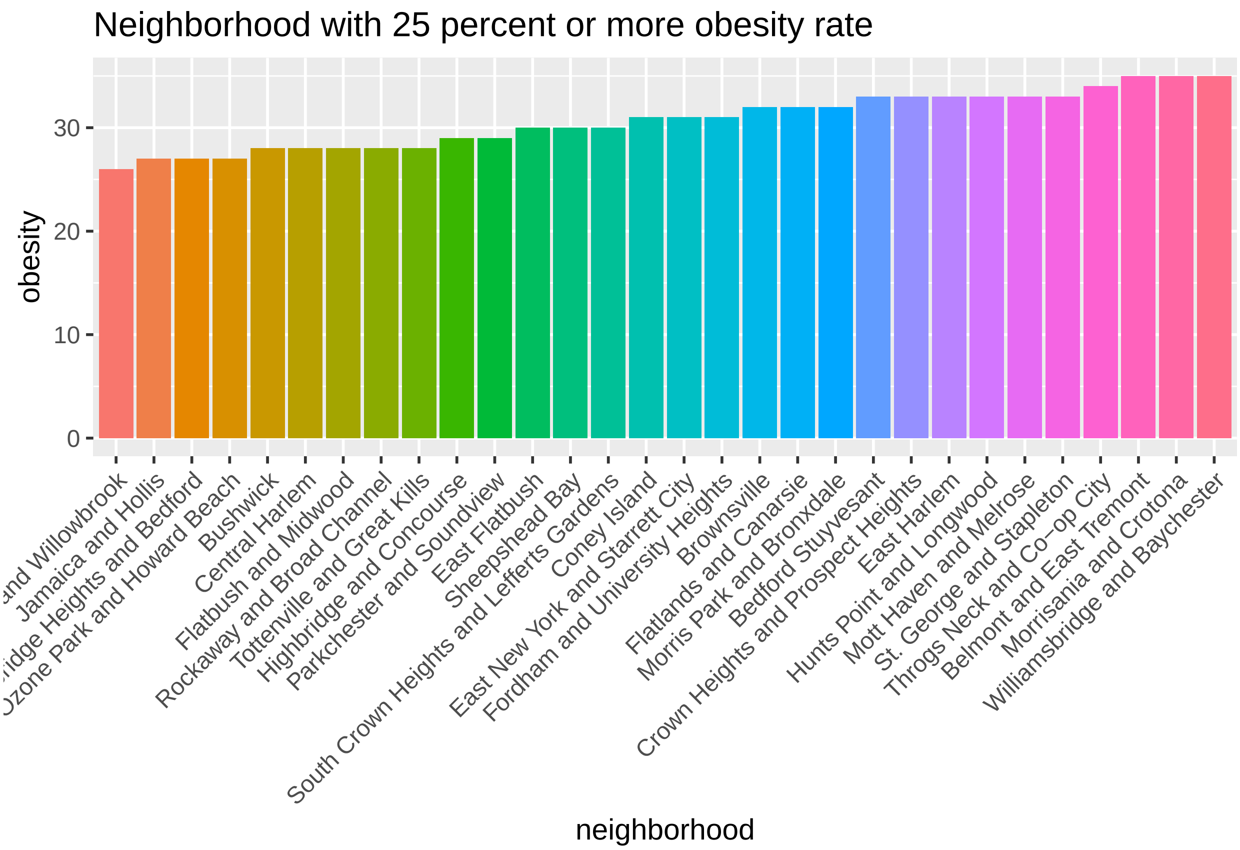
```r
zip_neighbor <- read_csv("neigh_zipcode.csv") %>%
  mutate(zipcode = as.character(zipcode))
##restaurant data with neighbourhood
rest_neighborhood = left_join(rest_inspection, zip_neighbor, by = "zipcode") %>%
  filter(!is.na(neighborhood))
```

Exploratory Analysis

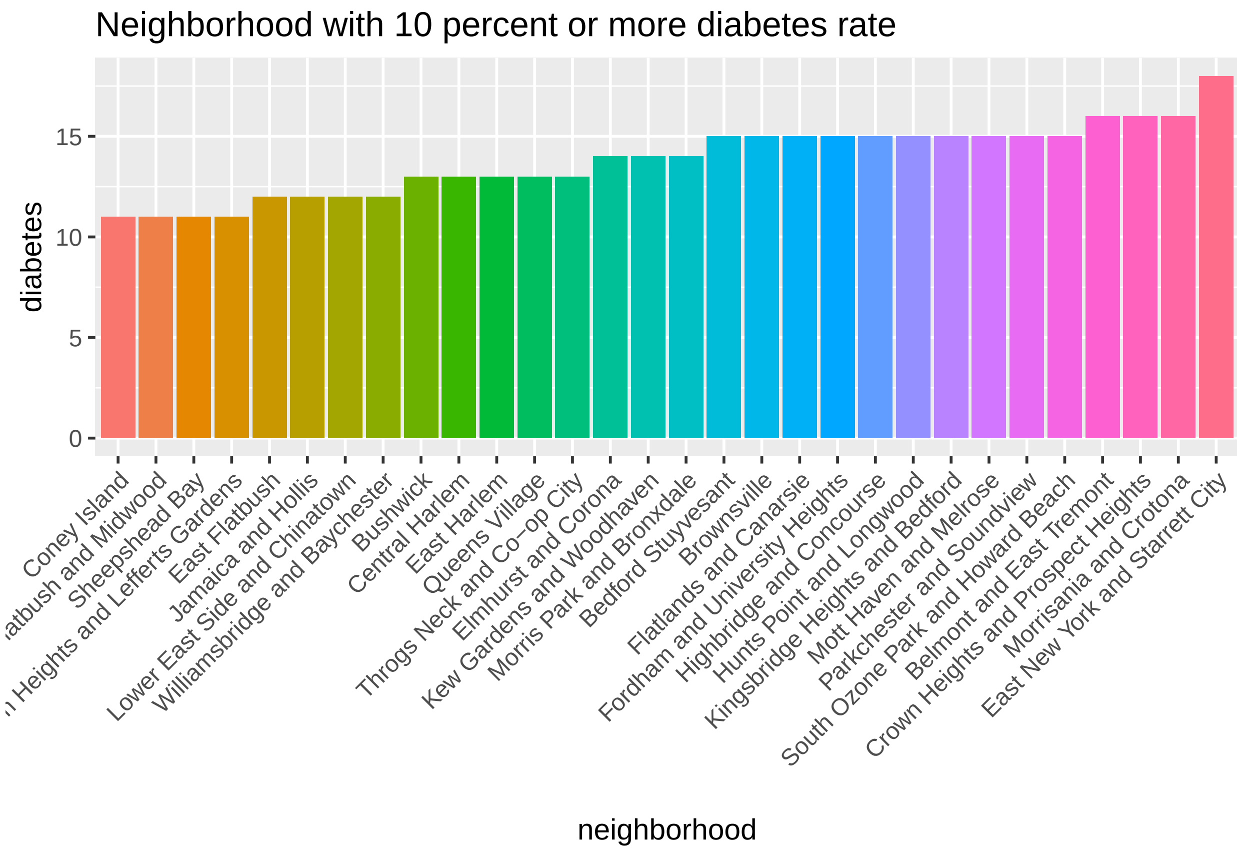
Health Data

The health data we achieved is very well-structured. We will make plots to see the distribution of the three chronic disease outcomes across neighorhoods.

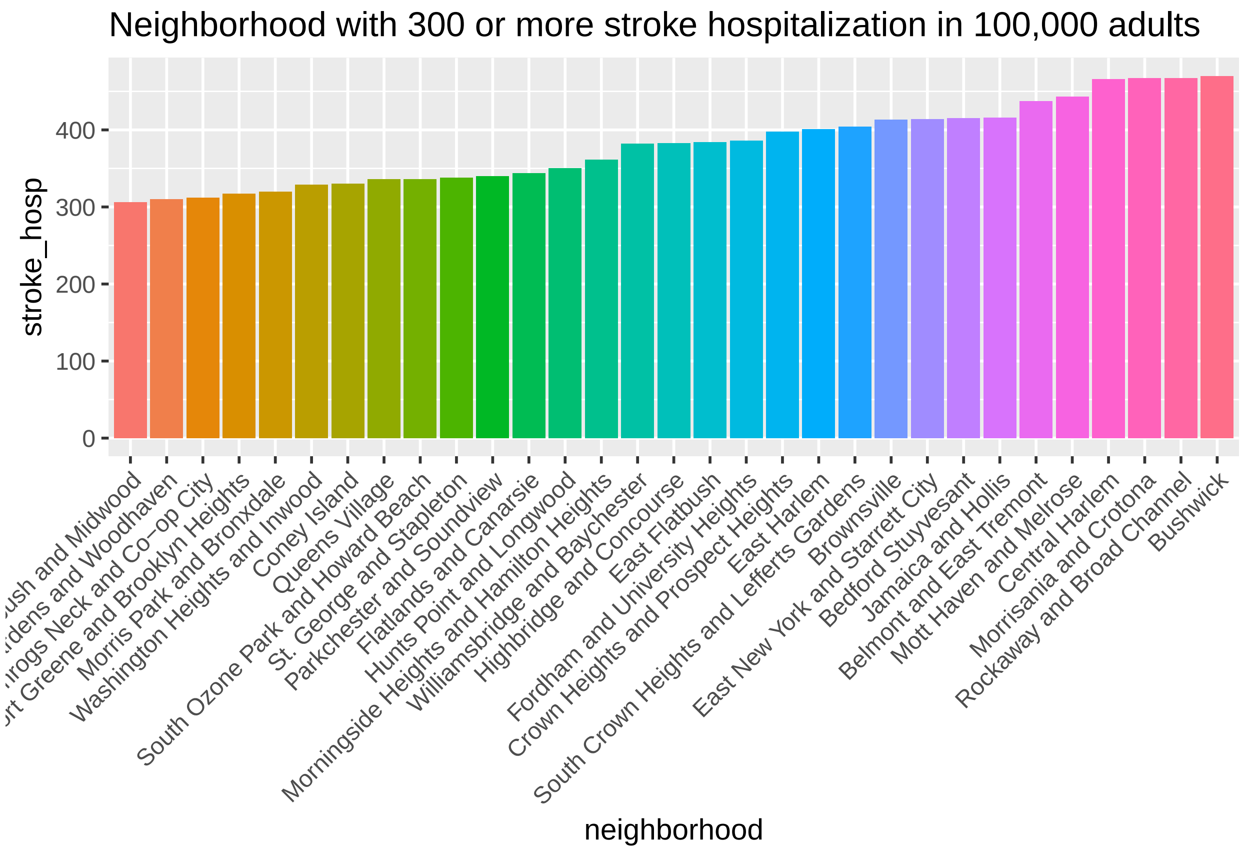
```r
health_only_neighborhood <- health[-c(1:6),] %>%
  rename(neighborhood = name) %>%
  mutate(neighborhood = as.factor(neighborhood))

##Plotting for outcome in different neighborhood
bar_obe <- health_only_neighborhood %>%
  mutate(neighborhood = fct_reorder(neighborhood, obesity)) %>%
  filter(obesity > 25) %>%
  ggplot(aes(x = neighborhood,y = obesity, fill = neighborhood)) + geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none") +
  labs(title = "Neighborhood with 25 percent or more obesity rate")
bar_dia <- health_only_neighborhood %>%
  mutate(neighborhood = fct_reorder(neighborhood, diabetes)) %>%
  filter(diabetes > 10) %>%
  ggplot(aes(x = neighborhood,y = diabetes, fill = neighborhood)) + geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none") +
  labs(title = "Neighborhood with 10 percent or more diabetes rate")
bar_stro <- health_only_neighborhood %>%
  mutate(neighborhood = fct_reorder(neighborhood, stroke_hosp)) %>%
  filter(stroke_hosp > 300) %>%
  ggplot(aes(x = neighborhood,y = stroke_hosp, fill = neighborhood)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none") +
  labs(title = "Neighborhood with 300 or more stroke hospitalization in 100,000 adults")
#ggplotly(bar_obe)
#ggplotly(bar_dia)
#ggplotly(bar_stro)
bar_obe
```


Neighborhood with 25 percent or more obesity rate
obesity
0
10
20
30
neighborhood
Jamaica and Hollis
Ozone Park and Howard Beach
Bushwick
Central Harlem
Flatbush and Midwood
Rockaway and Broad Channel
Tottenville and Great Kills
Highbridge and Concourse
Parkchester and Soundview
East Flatbush
Sheepshead Bay
South Crown Heights and Lefferts Gardens
Coney Island
East New York and Starrett City
Fordham and University Heights
Brownsville
Flatlands and Canarsie
Morris Park and Bronxdale
Bedford Stuyvesant
Crown Heights and Prospect Heights
East Harlem
Hunts Point and Longwood
Mott Haven and Melrose
St. George and Stapleton
Throgs Neck and Co−op City
Belmont and East Tremont
Morrisania and Crotona
Williamsbridge and Baychester

```
bar_dia
```


Neighborhood with 10 percent or more diabetes rate
diabetes
0
5
10
15
neighborhood
Coney Island
Sheepshead Bay
East Flatbush
Jamaica and Hollis
Lower East Side and Chinatown
Williamsbridge and Baychester
Bushwick
Central Harlem
East Harlem
Queens Village
Throgs Neck and Co−op City
Elmhurst and Corona
Kew Gardens and Woodhaven
Morris Park and Bronxdale
Bedford Stuyvesant
Brownsville
Flatlands and Canarsie
Fordham and University Heights
Highbridge and Concourse
Hunts Point and Longwood
Kingsbridge Heights and Bedford
Mott Haven and Melrose
Parkchester and Soundview
South Ozone Park and Howard Beach
Belmont and East Tremont
Crown Heights and Prospect Heights
Morrisania and Crotona
East New York and Starrett City

```
bar_stro
```



"'

Williamsbridge and Baychester along with two other neighborhoods (Belmont and East Tremont, Morrisania and Crotona) have the highest obesity rate, up to 35%. East New York and Starrett City have the highest diabetes rate, up to 18 percent. Bushwick has the highest stroke hospitalization in 100,000 adults, up to 470 adults.

Restaurant data

Fastfood restaurants by cuisine type

Fastfood restaurants are identified by their cuisine descriptions given in the inspection data. We print out the cuisine descriptions list (n=85) and let everyone circle the ones they think is fastfood and the union are used as our rule (use union because it's more conservative).

We classify cuisine descriptions "Bagels/Pretzels", "Bottled beverages, including water, sodas, juices, etc.", "Chicken", "Delicatessen", "Donuts", "Hamburgers", "Hotdogs", "Hotdogs/Pretzels", "Ice Cream, Gelato, Yogurt, Ices", "Nuts/Confectionary", "Pancakes/Waffles", "Pizza", "Soul Food", "Sandwiches", "Sandwiches/Salads/Mixed Buffet" and "Soups & Sandwiches" as fastfood restaurants. And then we calculate the total number of restaurants and the number of fastfood restaurants, as well as the percentage of fastfood restaurants for each neighborhood.

```
# calculating the total number of restaurants and the number of fastfood restaurants in the neighborhoo
neighborhood_list =
  rest_neighborhood %>%
  distinct(neighborhood) %>%
  arrange(neighborhood)
```

```r
rest_fastfood_neighborhood =
  rest_neighborhood %>%
  filter(cuisine_description %in% c("Bagels/Pretzels",
                                    "Bottled beverages, including water, sodas, juices, etc.",
                                    "Chicken",
                                    "Delicatessen",
                                    "Donuts",
                                    "Hamburgers",
                                    "Hotdogs",
                                    "Hotdogs/Pretzels",
                                    "Ice Cream, Gelato, Yogurt, Ices",
                                    "Nuts/Confectionary",
                                    "Pancakes/Waffles",
                                    "Pizza",
                                    "Soul Food",
                                    "Sandwiches",
                                    "Sandwiches/Salads/Mixed Buffet",
                                    "Soups & Sandwiches"))

percent_fastfood_neighborhood = function(name_neighborhood){

  rest_each_neighborhood =
    rest_neighborhood %>%
    filter(neighborhood == name_neighborhood) %>%
    distinct(camis)

  n_rest_neighborhood = nrow(rest_each_neighborhood)

  rest_fastfood_distinct_neighborhood =
    rest_fastfood_neighborhood %>%
    filter(neighborhood == name_neighborhood) %>%
    distinct(camis, cuisine_description)

  n_fastfood_neighborhood = nrow(rest_fastfood_distinct_neighborhood)

  percent_fastfood_neighborhood = n_fastfood_neighborhood/n_rest_neighborhood

  tibble(
    neighborhood = name_neighborhood,
    n_fastfood = n_fastfood_neighborhood,
    n_rest = n_rest_neighborhood,
    percent_fastfood = percent_fastfood_neighborhood
  )
}

fastfood_neighborhood =
  map(neighborhood_list$neighborhood, percent_fastfood_neighborhood) %>%
  bind_rows() %>%
  mutate(neighborhood = str_to_upper(neighborhood))

# plot for each neighborhood
# fastfood_neighborhood %>%
#   mutate(neighborhood = as.factor(neighborhood),
```

```
#           n_rest = as.numeric(n_rest),
#           n_nonfastfood = n_rest - n_fastfood,
#           neighborhood = fct_reorder(neighborhood, percent_fastfood)) %>%
#    plot_ly(., x = ~neighborhood, y = ~n_fastfood, type = 'bar', name = 'fastfood restaurants') %>%
#    add_trace(y = ~n_nonfastfood, name = 'non-fastfood restaurants') %>%
#    layout(yaxis = list(title = 'Number of restaurants'),
#           xaxis = list(title = 'Neighborhood (ordered by percentage of fastfood restaurants)',
#                        showticklabels = FALSE),
#           barmode = 'stack')
```

While the Greenwich Village and SOHO neighborhood has fairly large number of restaurants, it has the smallest percentage of fastfood restaurants. Williamsbridge and Baychester has the largest percentage of fastfood restaurants.

When large number of total restaurants is not equal to large percentage of fastfood restaurants in that neighborhood, we can conclude that the distribution of fastfood restaurants is not even across neighborhoods, which also implies the motivation of our study, we want to investigate if this uneven distribution of fastfood restaurants is associated with diffferent level of chronic disease outcomes within a neighborhood.

Restaurant Chains

We first scrape the list of 75 national chain restaurants in the US from the wikipedia page (https://en.wikipedia.org/wiki/List_of_restaurant_chains_in_the_United_States#Fast-casual) and then join this dataset with restaurant inspection data to choose only the chain restaurants in NYC.

```
chains_html = read_html("https://en.wikipedia.org/wiki/List_of_restaurant_chains_in_the_United_States#Fa

# read in the list of chain restaurants in us
# made the names to uppercase and changed the var name to dba
chain_rest = chains_html %>%
  html_nodes("td:nth-child(1)") %>%
  html_text() %>%
  as.tibble() %>%
  mutate(dba = value,
         dba = str_to_upper(dba),
         dba = str_replace_all(dba, "[\r\n]" , "")) %>%
  select(dba)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
head(chain_rest, 10)
```

```
## # A tibble: 10 x 1
##    dba
##    <chr>
##  1 ""
##  2 A&W RESTAURANTS
##  3 AMERICA'S INCREDIBLE PIZZA COMPANY
##  4 APPLEBEE'S
##  5 ARBY'S
##  6 ARCTIC CIRCLE RESTAURANTS
##  7 ARTHUR TREACHER'S
##  8 ATLANTA BREAD COMPANY
##  9 AUNTIE ANNE'S
```

```
## 10 BAHAMA BREEZE
```

Then, we match the list of chain restaurants in U.S. with the restaurant inspection data.

```r
# removing punctuations in chain_rest & restaurant inspections (neighborhoods)
chain_rest_str =
  chain_rest  %>%
  mutate(dba = str_replace_all(dba, "[[:punct:]]", ""))

rest_neigh_str = rest_neighborhood %>%
  mutate(dba = str_replace_all(dba, "[[:punct:]]", ""))

# Matching the two datasets(restaurant inspection data that has all punctuation removed from dba(restau
neighborhood_chain =
  right_join(rest_neigh_str, chain_rest_str) %>%
  filter(!is.na(camis)) %>%
  distinct(camis, dba, neighborhood, boro)
```

```
## Joining, by = "dba"
```

```r
neighborhood_chain %>%
  group_by(dba) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    dba                         n
##    <chr>                   <int>
##  1 STARBUCKS                 286
##  2 SUBWAY                    285
##  3 MCDONALDS                 205
##  4 POPEYES                   101
##  5 BURGER KING                78
##  6 CHIPOTLE MEXICAN GRILL     76
##  7 DUNKIN DONUTS              51
##  8 WENDYS                     43
##  9 KFC                        36
## 10 LITTLE CAESARS             35
```

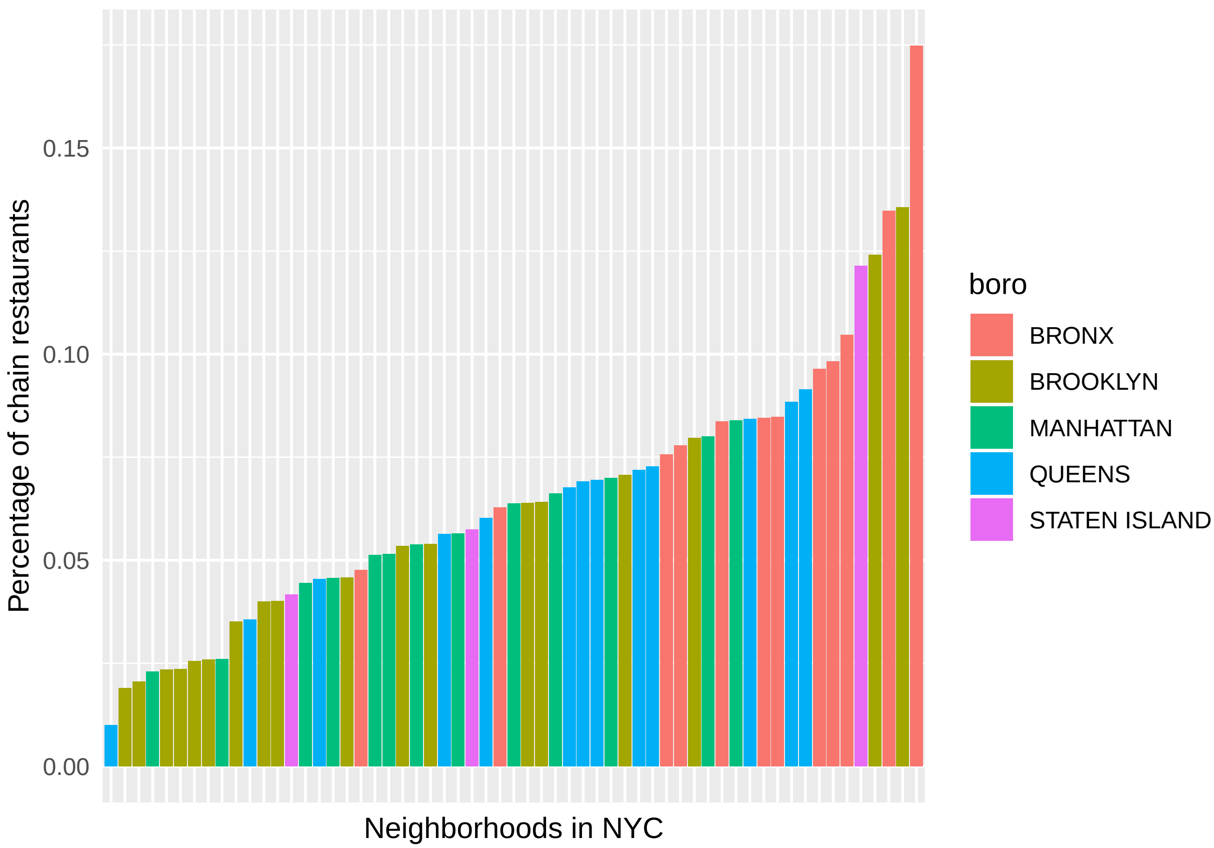
The combined dataset "neighborhood_chain" has 1546 observations. Also, there were 64 different chain restaurants extracted.

```r
# counting chains in neighborhoods
neigh_count_chain = neighborhood_chain %>%
  group_by(neighborhood, boro) %>%
  summarise(chain_n = n())

neigh_count_rest = rest_neighborhood %>%
  distinct(neighborhood, camis) %>%
  group_by(neighborhood) %>%
  summarise(res_n = n())

# calculating percentage of chains in each neighborhood
percent_neighborhood_chain = left_join(neigh_count_chain, neigh_count_rest) %>%
  ungroup() %>%
  mutate(chain_percentage = chain_n/res_n,
```

```
         neighborhood = str_to_upper(neighborhood))
```

```
## Joining, by = "neighborhood"
```

```
plot_chain_neighbor = percent_neighborhood_chain %>%
  mutate(neighborhood = forcats::fct_reorder(neighborhood, chain_percentage)) %>%
  ggplot(aes(neighborhood, chain_percentage, fill = boro)) + geom_bar(stat="identity") +
  labs(x = "Neighborhoods in NYC", y = "Percentage of chain restaurants") +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank())

#ggplotly(plot_chain_neighbor)
plot_chain_neighbor
```




```
max(percent_neighborhood_chain$chain_percentage)
```

```
## [1] 0.1748252
```

We plot neighborhoods in NYC with their percentage of chain restaurants and group them by borough. We can see that neighborhoods with those smallest percentages of chain restuarants are mostly in Brooklyn except for Greenwhich Village and Soho and Lower East Side and Chinatown in Manhattan. Neighborhoods in Queens and Manhattan are spread out across low to high percentage of chain restaurants while most of the neighborhoods in Bronx and Staten Island have high percentages. The neighborhood with the highest percentage of chain restaurants is Throgs neck and Co-op City in Bronx with 17.5% of chain restaurats out of all restaurats.

Inspection Grade

```
gradea_neighborhood =
  rest_neighborhood %>%
  group_by(neighborhood, grade) %>%
  summarise(n = n()) %>%
  mutate(grade_percent = n / sum(n)) %>%
  filter(grade == "A") %>%
  ungroup(boro) %>%
  mutate(neighborhood = str_to_upper(neighborhood))

# gradea_neighborhood %>%
#   mutate(neighborhood = as.factor(neighborhood),
#          neighborhood = fct_reorder(neighborhood,grade_percent)) %>%
#   plot_ly(x = ~neighborhood, y = ~grade_percent, color = ~neighborhood, type = "bar") %>%
#   layout(yaxis = list(title = 'Percentage of Grade A restaurants'),
#          xaxis = list(title = 'Neighborhoods in NYC', showticklabels = FALSE),
#          showlegend = FALSE)

gradea_rank = gradea_neighborhood %>%
  mutate(neighborhood = as.factor(neighborhood),
         neighborhood = fct_reorder(neighborhood,grade_percent)) %>%
         arrange(desc(grade_percent))
```

The differences on percentage of "grade-A" restaurants between each neighborhood are observed. "Tottenville and Great Kills" has the greatest grade-A restaurant percentage, around 50.7%. "Sunset Park", however, has the least, around 32.4%. "Washington Heights", where we live, takes the second to last, around 32%, which is obviously consistent with the feeling we have towards the restaurant condition of "Washington Heights"

Formal Analysis and Findings

Model Selection Process

First, we match the datasets containing all the restaurant information with the health datasets by common variable "neighborhood".

```
health_neighborhood =
  health %>%
  mutate(neighborhood = str_to_upper(name)) %>%
  select(-name)

combined_chain =
  percent_neighborhood_chain %>%
  select(neighborhood, chain_percentage)
combined_chain_fastfood =
  fastfood_neighborhood %>%
  mutate(fastfood_percent = percent_fastfood) %>%
  select(neighborhood, fastfood_percent) %>%
  right_join(combined_chain, by = "neighborhood")
combined_chain_fastfood_gradea =
  gradea_neighborhood %>%
  select(neighborhood, grade_percent) %>%
```

```
  right_join(combined_chain_fastfood, by = "neighborhood")

combined_model =
  left_join(combined_chain_fastfood_gradea, health_neighborhood, by = "neighborhood")
```

Main predictor selection

```
outcome_name = combined_model[,10:12]

main_predictor_selection = function(outcome){

lm_fastfood =
  lm(outcome ~ fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise + airqua
  summary()
lm_fastfood_tbl =
  as.tibble(lm_fastfood[[4]]) %>%
  clean_names()

lm_chain =
  lm(outcome ~ chain_percentage + grade_percent + racewhite_rate + poverty + smoking + exercise + airqua
  summary()
lm_chain_tbl =
  as.tibble(lm_chain[[4]]) %>%
  clean_names()

lm_both =
  lm(outcome ~ fastfood_percent + chain_percentage + grade_percent + racewhite_rate + poverty + smoking
  summary()
lm_both_tbl =
  as.tibble(lm_both[[4]]) %>%
  clean_names()

tibble(p_fastfood_sing = lm_fastfood_tbl$pr_t[2],
       p_chain_sing = lm_chain_tbl$pr_t[2],
       p_fastfood_both = lm_both_tbl$pr_t[2],
       p_chain_both = lm_both_tbl$pr_t[3])
}

main_predictor_comp =
  map(outcome_name, main_predictor_selection) %>%
  bind_rows()
main_predictor_comp$outcome = colnames(combined_model)[10:12]
main_predictor_comp
```

```
## # A tibble: 3 x 5
##   p_fastfood_sing p_chain_sing p_fastfood_both p_chain_both outcome
##             <dbl>        <dbl>           <dbl>        <dbl> <chr>
## 1        0.000305      0.00221          0.0370        0.387 obesity
## 2        0.0313        0.0609           0.239         0.594 diabetes
## 3        0.0137        0.340            0.0126        0.280 stroke_hosp
```

We originally have two main predictors of interest, *Percentage of chain restaurants* and *Percentage of fastfood restaurants*, and they are both significantly associated with the three chronic disease health oucomes when solely in the model after adjusting for other potential confounders. Taking obesity as example, the p-value for fastfood_percent is 0.00030506 and for chain_percentage is 0.0022107.

And we also anticipated them to be highly associated with each other, to avoid collinearity, we need to make decision on which one to keep as the final main predictor. So we put the two main predictor candidates in the same model and see how their p-values change. As a result, the fastfood_percent stays significant (p=0.0370412) and the chain_percentage turn insignificant (p=0.3867866).

Same selecting processes are repeated for the oucomes diabetes and stroke. Thus, we decide to have *Percentage of fastfood restaurants* (fastfood_percent) as the main predictor.

Confounder selection

```
outcome_name = combined_model[,10:12]

confounder_percent_change = function(outcome){
lm_adjusted =
  lm(outcome ~ fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise + airqua
  summary()
lm_adjusted_tbl =
  as.tibble(lm_adjusted[[4]]) %>%
  clean_names()

lm_crude =
  lm(outcome ~ fastfood_percent + racewhite_rate + poverty + smoking + exercise + airquality_rate, combi
  summary()
lm_crude_tbl =
  as.tibble(lm_crude[[4]]) %>%
  clean_names()

percent_change = (lm_crude_tbl$estimate[2] - lm_adjusted_tbl$estimate[2]) /lm_crude_tbl$estimate[2]

confounder_percent_change = percent_change
}

confounder_change =
  map(outcome_name, confounder_percent_change) %>%
  bind_rows()

percent <- function(x) {
  paste0(formatC(100 * x, digits = 3), "%")
}

confounder_change %>%
  kable()
```

obesity	diabetes	stroke_hosp
-0.0921697	0.2017842	-0.111628

Besides some biologically meaningful covariates (i.e. race, poverty, smoking status, exercise), we also hypothesize the variable *Percentage of grade A restuarants* as a potential confounder in the association between fastfood restaurants percentage and the three chronic disease health outcomes.

Here, we are assessing if grade_percent is a significant confounder. Using the 10% change rule of thumb, we find that after adjusting for *Percentage of grade A restuarants*, the estimates of fastfood_percent change by -9.22% for outcome obsesity, 20.2% for outcome diabetes and -11.2% for outcome stroke hospitalization.

Thus, regarding the final models, we are going to keep grade_percent for obsesity and stroke hospitalization, but take it out and rerun the model for the outcome diabetes.

Final Models and Findings

Model 1: Obesity = fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise

```
lm_obesity =
  lm(obesity ~ fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise+ airqual
  broom::tidy()
lm_obesity %>%
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	22.1249808	12.9762514	1.7050364	0.0942727
fastfood_percent	50.9269225	13.1405393	3.8755580	0.0003051
grade_percent	-12.2432962	18.7914498	-0.6515355	0.5176262
racewhite_rate	-0.0542501	0.0377006	-1.4389706	0.1562683
poverty	0.2266346	0.0955312	2.3723619	0.0214847
smoking	0.6263512	0.2162401	2.8965538	0.0055450
exercise	0.1008898	0.1565246	0.6445623	0.5220996
airquality_rate	-2.5297920	0.6365427	-3.9742694	0.0002228

Model 2: Diabetes = fastfood_percent + racewhite_rate + poverty + smoking + exercise

```
lm_diabetes =
  lm(diabetes ~ fastfood_percent + racewhite_rate + poverty + smoking + exercise, combined_model) %>%
  broom::tidy()
lm_diabetes %>%
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.6193246	5.2290090	3.560775	0.0007906
fastfood_percent	18.2372209	5.3491925	3.409341	0.0012518
racewhite_rate	-0.0744485	0.0167050	-4.456648	0.0000435
poverty	0.0407262	0.0376155	1.082697	0.2838441
smoking	0.1497904	0.1006810	1.487772	0.1427370
exercise	-0.1571246	0.0585783	-2.682300	0.0097290

Model 3: Stroke_hosp = fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise

```
lm_stroke =
  lm(stroke_hosp ~ fastfood_percent + grade_percent + racewhite_rate + poverty + smoking + exercise, co
  broom::tidy()
lm_stroke %>%
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	87.356938	157.2876781	0.5553959	0.5810055
fastfood_percent	441.282120	160.3406884	2.7521531	0.0081321
grade_percent	-131.585238	231.2544086	-0.5690064	0.5718014
racewhite_rate	-1.517473	0.4517671	-3.3589711	0.0014706
poverty	1.529170	1.0218840	1.4964221	0.1405890
smoking	6.824540	2.6291733	2.5956981	0.0122417
exercise	1.395995	1.5194339	0.9187597	0.3624635

We have three final models each having a health outcome (prevalence of obesity, prevalence of diabetes, and stroke hospitalization rate) as the dependent variable and percentage of fastfood as the main predictor. Obesity and percentage of fastfood restuarants model gives the most significant results with p-value of the main predictor being 0.00030506 (<0.01). In other words, at significance level of 1%, every 10% increase in the number of fastfood restaurants (0.1 unit increase in the percentage of fastfood restaurants) in a neighborhood is associated with 5.09% increase in the neighborhood's obesity prevalence, while adjusting for other factors.

The p-value of the regression coefficients for model with outcomes as diabetes/stroke are both less than 0.01 (Diabetes p=0.0012518; Stroke p=0.0081321), indicating there is a significant linear relationship between diabetese/stroke and percentage of fastfood restaurants at 1% significance level. To put in other words, at significance level of 1%, every 10% increase in the number of fastfood restaurants in a neighborhood is associated with 1.82% increase in the neighborhood's diabetes prevalence, while adjusting for other factors. Furthermore, every 10% increase in the number of fastfood restaurants in a neighborhood is associated with increase in 44.1 stroke hospitalizations per 100,000 adults in the neighborhood, while adjusting for other factors.

It is reasonable that among the three health outcomes, obesity has the strongest linear association with food environment, in this case, percentage of fastfood restaurants. It is a health condition that has the most direct relation with one's diet pattern. Moreover, it has a higher prevalence than diabetes and stroke, which could lead to a lower p-value than the other two health conditions.

Conclusion

Overall, we conclude that there is a significant relationship between chronic disease outcomes (i.e. obesity, diabetes, stroke) and the geographical distribution of fast-food restaurants in New York City.

Discussion

Our 2015 Community Health Profiles Open Data has data on obesity, diabetes prevalence rate from 2013 and stroke hospitalization rate from 2012. And the data of geographic distribution of restaurants is from 2014 to 2019. So the health profile data is somewhat earlier than the restaurant geographic distribution. However, as restaurant distribution and chronic disease prevalence rate wouldn't change much over a few years, the lag between years is not a large concern and theoretically won't affect our results that much. Therefore, our results can still be valid.

Although the cuisine type of a restaurant is typically assumed to directly reflect the health level of the food that restaurant provides, we demonstrate here that even the unhealthiest restaurant can offer one or several healthy foods, which may bring bias to our research. Besides, our study at first treats chain restaurants as a criteria to identify unhealthy restaurants. Obviously this criteria is ambiguous and farfetched. In the end, the study doesn't include this as the predictor.

Lastly, although there is an association between fast-food restaurant geographical distribution and chronic disease outcomes, we can not conclude that it is fast-food restaurant causing these health outcomes. Most

People do not eat out all the time and we cannot to assume the prevalent restaurant types in a region identifies the residents' diet. Further study need to be conducted before concluding any causation effect.