# Relationship Between Number of Fast Food Chain and Diseases

COL NYC 2019: ETL Project

Due: 10/19/2019

By: Pedro Takenouchi, Jorge Sanchez, Kwang Min Ko

## Executive Summary

Our project is inspired by movie, “Super Size Me”, we want to find out the relationship between numbers of fast food chain impacting the health and diseases of the population in the chosen area. We have identified the list of dies five most common physical and mental illness that can be caused by consistently consuming the fast foods. This is the document describing the ETL process that we carried out using the two data we found; “Fast Food Restaurants Across America”, and “Health search by US Metropolitan Area, 2005-2017”.

## Data Dictionary

| Name | Description | Type | Value | Unique |
|---|---|---|---|---|
| Province | States (ex, NY) | Varchar | Not Null | Yes |
| City | City (ex, New Jersey) | Varchar | Not Null | Yes |
| Name | Name of the Fast Food Chain Restaurant | Varchar | Not Null | Yes |
| dma | States, City | Varchar | Not Null | No |
| 2017+Diarrhea | Diseases Type 1 | Integer | Not Null | Yes |
| 2017+Obesity | Diseases TYPE 2 | Integer | Not Null | Yes |
| 2017+Diabetes | Diseases TYPE 3 | Integer | Not Null | Yes |
| v2017+Cancer | Diseases TYPE 4 | Integer | Not Null | Yes |
| 2017+Depression | Diseases TYPE 5 | Integer | Not Null | Yes |
| 2016+Diarrhea | Diseases Type 1.2 | Integer | Not Null | Yes |
| 2016+Obesity | Diseases TYPE 2.2 | Integer | Not Null | Yes |
| 2016+Diabetes | Diseases TYPE 3.3 | Integer | Not Null | Yes |
| 2016+Cancer | Diseases TYPE 4.2 | Integer | Not Null | Yes |
| 2016+Depression | Diseases TYPE 5.2 | Integer | Not Null | Yes |
| 2015+Diarrhea | Diseases Type 1.3 | Integer | Not Null | Yes |
| 2015+Obesity | Diseases TYPE 2.3 | Integer | Not Null | Yes |
| 2015+Diabetes | Diseases TYPE 3.3 | Integer | Not Null | Yes |
| 2015+Cancer | Diseases TYPE 4.3 | Integer | Not Null | Yes |
| 2015+Depression | Diseases TYPE 5.3 | Integer | Not Null | Yes |

# Data Cleaning

## Part 1: Fast Food Restaurant Data Cleaning

We cleaned the data by taking out each fast food restaurant, city and state data. We made graph to see which state and city has most fast food restaurant numbers.

**Transform FastFood DataFrame**

In [30]:
```
# Create a filtered dataframe from specific columns
fastfood_cols = ["address", "city", "country","keys","latitude","longitude","name","postalCode","province","websites"]
fastfood_transformed= fastfood_df[fastfood_cols].copy()

# Rename the column headers
fastfood_transformed = fastfood_transformed.rename(columns={"keys": "id","province": "state"})

# Clean the data by dropping duplicates and setting the index
fastfood_transformed.drop_duplicates("id", inplace=True)
fastfood_transformed.set_index("id", inplace=True)

fastfood_cleaned_cols = ["name", "city", "state"]
fastfood_cleaned= fastfood_transformed[fastfood_cleaned_cols].copy()
fastfood_cleaned
```

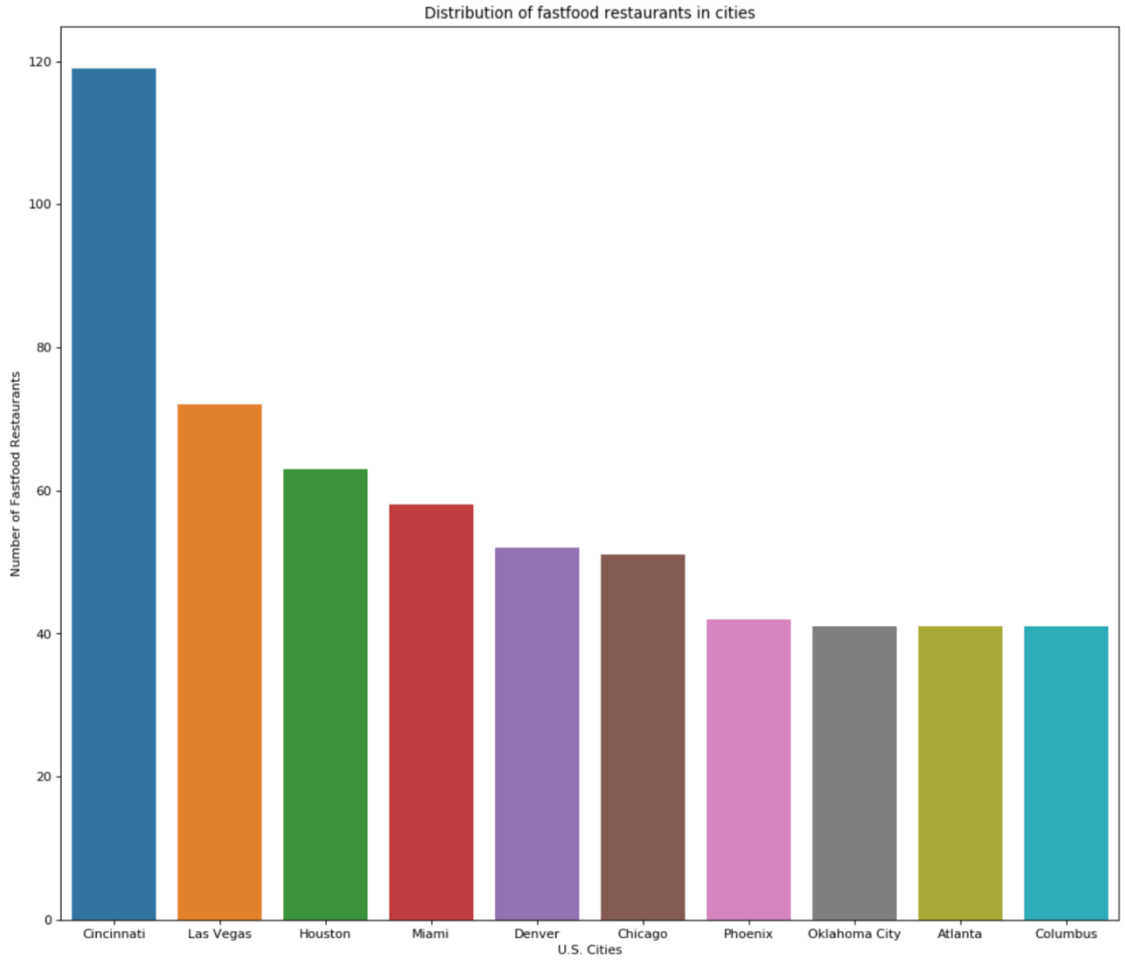
Out[30]:

| id | name | city | state |
|---|---|---|---|
| us/ny/massena/324mainst/-1161002137 | McDonald's | Massena | NY |
| us/oh/washingtoncourthouse/530clintonave/-791445730 | Wendy's | Washington Court House | OH |
| us/ky/maysville/408marketsquaredr/1051460804 | Frisch's Big Boy | Maysville | KY |
| us/ny/massena/6098statehighway37/-1161002137 | McDonald's | Massena | NY |
| us/oh/athens/139columbusrd/990890980 | OMG! Rotisserie | Athens | OH |
| us/oh/hamilton/4182tonyatrl/-1055723171 | Domino's Pizza | Hamilton | OH |
| us/oh/englewood/590smainst/-1055723171 | Domino's Pizza | Englewood | OH |
| us/sc/saluda/401njenningsst/-1161002137 | McDonald's | Saluda | SC |
| us/sc/batesburg/205wchurchst/-791445730 | Wendy's | Batesburg | SC |

Cincinnati and Las Vegas have most number of Fast Food Restaurants in States.

In [64]:
```
#Distribution of fastfood restaurants in US cities

figure(num=None, figsize=(15, 13), dpi=80, facecolor='w', edgecolor='k')
sns.barplot(x=fastfood_df.city.value_counts().index, y=fastfood_df.city.value_counts(),
          order = fastfood_df.city.value_counts().iloc[:10].index)

plt.xlabel("U.S. Cities")
plt.ylabel("Number of Fastfood Restaurants")
plt.title("Distribution of fastfood restaurants in cities")

#We can see that the metropolitan area with the most fast food restaurants
# is Cincinatti,
```

```
Out[64]: Text(0.5, 1.0, 'Distribution of fastfood restaurants in cities')
```




## Part 2: Diseases Data Cleaning

We have chosen top five diseases that could be caused consistently consuming fast food. We decided five diseases based on our research and graph we have created.

**Transform Metropolitan DataFrame**

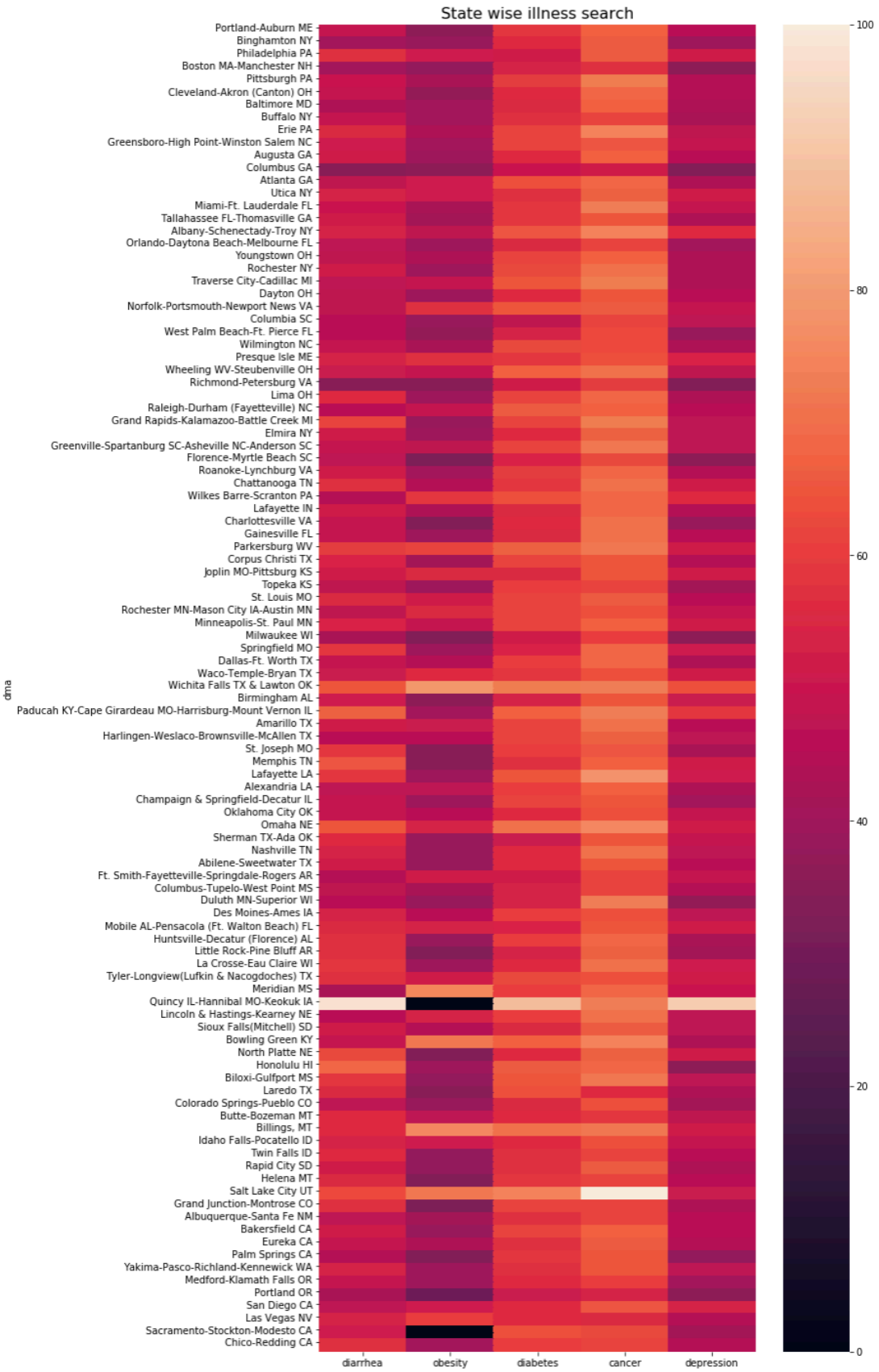
```
In [31]: # Create a filtered dataframe from specific columns, only get 2015, 2016, and 2017 data
         # There are no nulls or duplicates in this dataset

         metro_cols = ["dma", "2015+diarrhea", "2015+obesity", "2015+diabetes", "2015+cancer", "2015+depression", "2016+diarrhea
         metro_transformed= metropolitan_df[metro_cols].copy()
         metro_transformed
```

Out[31]:

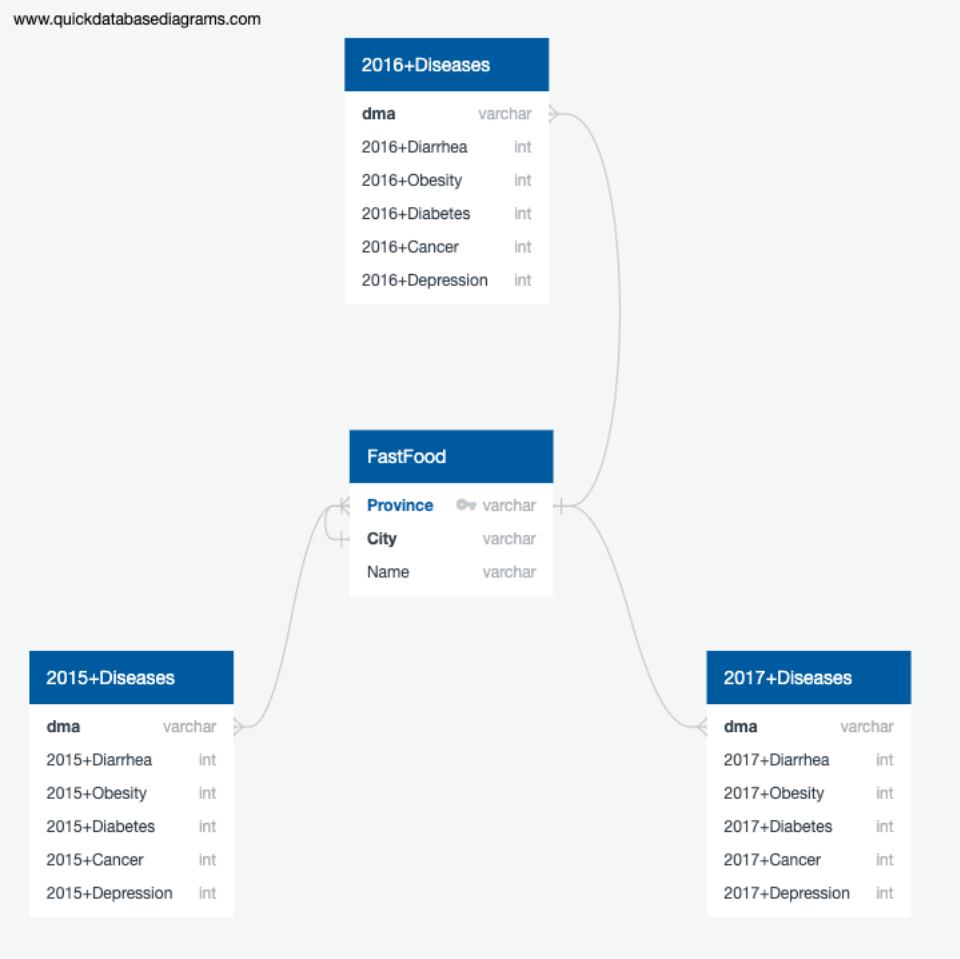
| | dma | 2015+diarrhea | 2015+obesity | 2015+diabetes | 2015+cancer | 2015+depression | 2016+diarrhea | 2016+obesity | 2016+diabetes | 2( |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Portland-Auburn ME | 68 | 48 | 73 | 67 | 70 | 69 | 49 | 81 | |
| 1 | New York NY | 56 | 46 | 63 | 64 | 55 | 57 | 49 | 77 | |
| 2 | Binghamton NY | 84 | 67 | 75 | 64 | 69 | 79 | 70 | 74 | |
| 3 | Macon GA | 71 | 60 | 73 | 66 | 68 | 66 | 51 | 78 | |
| 4 | Philadelphia PA | 67 | 56 | 72 | 71 | 63 | 70 | 52 | 80 | |
| 5 | Detroit MI | 69 | 46 | 65 | 67 | 61 | 71 | 43 | 73 | |
| 6 | Boston MA-Manchester NH | 62 | 58 | 71 | 69 | 64 | 63 | 53 | 76 | |

We created graph which location has highest illness, we found out Las Vegas the second most number of fast food chain restaurant has one of highest rate of illness.

In [62]:
```
statesData = pd.DataFrame(metro_transformed.iloc[:,0])
healthSearchData = metro_transformed.drop(['dma'],axis=1)

meanDict = {}
yearList = []
illnessList = []
for col in healthSearchData.columns:
    if '+' in col:
        yearList.append(col.split('+')[0])
        illnessList.append(col.split('+')[-1])

for index, row in healthSearchData.iterrows():
    for illness in illnessList:
        searchCountList = []
        for year in yearList:
            searchCountList.append(row[year+ '+' +illness])
        if not illness in meanDict:
            meanDict[illness] = []
        meanDict[illness].append(np.mean(searchCountList))
yearWiseMeanDf = pd.DataFrame.from_dict(meanDict, orient='columns', dtype=None)
heatMapData = statesData.join(yearWiseMeanDf)
heatMapData.set_index('dma', inplace=True, drop=True)

import seaborn as sns
plt.figure(figsize=(10, 25))
plt.title("State wise illness search", fontsize=16)
ax = plt.subplot(111)
ax.spines["top"].set_visible(False)
ax.spines["bottom"].set_visible(False)
ax.spines["right"].set_visible(False)
ax.spines["left"].set_visible(False)
ax.get_xaxis().tick_bottom()
ax.get_yaxis().tick_left()
ax = sns.heatmap(heatMapData)
```


State wise illness search
dma
Portland-Auburn ME
Binghamton NY
Philadelphia PA
Boston MA-Manchester NH
Pittsburgh PA
Cleveland-Akron (Canton) OH
Baltimore MD
Buffalo NY
Erie PA
Greensboro-High Point-Winston Salem NC
Augusta GA
Columbus GA
Atlanta GA
Utica NY
Miami-Ft. Lauderdale FL
Tallahassee FL-Thomasville GA
Albany-Schenectady-Troy NY
Orlando-Daytona Beach-Melbourne FL
Youngstown OH
Rochester NY
Traverse City-Cadillac MI
Dayton OH
Norfolk-Portsmouth-Newport News VA
Columbia SC
West Palm Beach-Ft. Pierce FL
Wilmington NC
Presque Isle ME
Wheeling WV-Steubenville OH
Richmond-Petersburg VA
Lima OH
Raleigh-Durham (Fayetteville) NC
Grand Rapids-Kalamazoo-Battle Creek MI
Elmira NY
Greenville-Spartanburg SC-Asheville NC-Anderson SC
Florence-Myrtle Beach SC
Roanoke-Lynchburg VA
Chattanooga TN
Wilkes Barre-Scranton PA
Lafayette IN
Charlottesville VA
Gainesville FL
Parkersburg WV
Corpus Christi TX
Joplin MO-Pittsburg KS
Topeka KS
St. Louis MO
Rochester MN-Mason City IA-Austin MN
Minneapolis-St. Paul MN
Milwaukee WI
Springfield MO
Dallas-Ft. Worth TX
Waco-Temple-Bryan TX
Wichita Falls TX & Lawton OK
Birmingham AL
Paducah KY-Cape Girardeau MO-Harrisburg-Mount Vernon IL
Amarillo TX
Harlingen-Weslaco-Brownsville-McAllen TX
St. Joseph MO
Memphis TN
Lafayette LA
Alexandria LA
Champaign & Springfield-Decatur IL
Oklahoma City OK
Omaha NE
Sherman TX-Ada OK
Nashville TN
Abilene-Sweetwater TX
Ft. Smith-Fayetteville-Springdale-Rogers AR
Columbus-Tupelo-West Point MS
Duluth MN-Superior WI
Des Moines-Ames IA
Mobile AL-Pensacola (Ft. Walton Beach) FL
Huntsville-Decatur (Florence) AL
Little Rock-Pine Bluff AR
La Crosse-Eau Claire WI
Tyler-Longview(Lufkin & Nacogdoches) TX
Meridian MS
Quincy IL-Hannibal MO-Keokuk IA
Lincoln & Hastings-Kearney NE
Sioux Falls(Mitchell) SD
Bowling Green KY
North Platte NE
Honolulu HI
Biloxi-Gulfport MS
Laredo TX
Colorado Springs-Pueblo CO
Butte-Bozeman MT
Billings, MT
Idaho Falls-Pocatello ID
Twin Falls ID
Rapid City SD
Helena MT
Salt Lake City UT
Grand Junction-Montrose CO
Albuquerque-Santa Fe NM
Bakersfield CA
Eureka CA
Palm Springs CA
Yakima-Pasco-Richland-Kennewick WA
Medford-Klamath Falls OR
Portland OR
San Diego CA
Las Vegas NV
Sacramento-Stockton-Modesto CA
Chico-Redding CA
diarrhea
obesity
diabetes
cancer
depression
100
80
60
40
20
0

# ERD Diagram


www.quickdatabasediagrams.com
2016+Diseases
dma varchar
2016+Diarrhea int
2016+Obesity int
2016+Diabetes int
2016+Cancer int
2016+Depression int
FastFood
Province varchar
City varchar
Name varchar
2015+Diseases
dma varchar
2015+Diarrhea int
2015+Obesity int
2015+Diabetes int
2015+Cancer int
2015+Depression int
2017+Diseases
dma varchar
2017+Diarrhea int
2017+Obesity int
2017+Diabetes int
2017+Cancer int
2017+Depression int

# Table Schema

```
FastFood
-
Province varchar FK >- FastFood.City
City varchar
Name varchar
Distribution int

Diseases
-
dma varchar pk FK >- FastFood.Province
2016+Diarrhea int
2016+Obesity int
2016+Diabetes int
2016+Cancer int
2016+Depression int
2017+Diarrhea int
2017+Obesity int
2017+Diabetes int
2017+Cancer int
2017+Depression int
```

# Queries

```
CREATE TABLE FastFood (
   Keys varchar  NOT NULL ,
   Province varchar  NOT NULL ,
   City varchar  NOT NULL ,
   Brand varchar  NOT NULL ,
   PRIMARY KEY (
       Keys
   )
);

CREATE TABLE 2017+Diseases (
   dma varchar  NOT NULL ,
   2017+Diarrhea int  NOT NULL ,
   2017+Obesity int  NOT NULL ,
   2017+Diabetes int  NOT NULL ,
   2017+Cancer int  NOT NULL ,
   2017+Depression int  NOT NULL
);
CREATE TABLE 2016+Diseases (
   dma varchar  NOT NULL ,
   2016+Diarrhea int  NOT NULL ,
   2016+Obesity int  NOT NULL ,
   2016+Diabetes int  NOT NULL ,
   2016+Cancer int  NOT NULL ,
   2016+Depression int  NOT NULL
);
CREATE TABLE 2015+Diseases (
   dma varchar  NOT NULL ,
   2015+Diarrhea int  NOT NULL ,
   2015+Obesity int  NOT NULL ,
   2015+Diabetes int  NOT NULL ,
   2015+Cancer int  NOT NULL ,
   2015+Depression int  NOT NULL

TRUNCATE TABLE FastFood
```