Diedre Brown

INFO 640/Fall 2019 Data Analysis

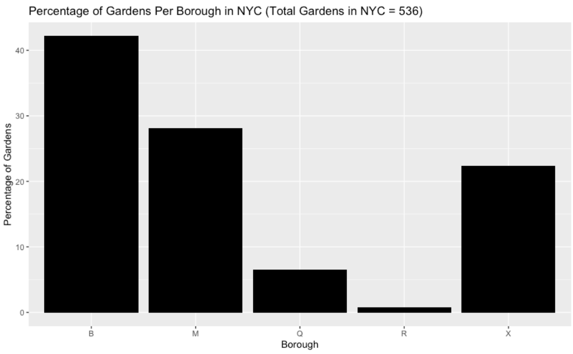
Descriptive Data Analysis

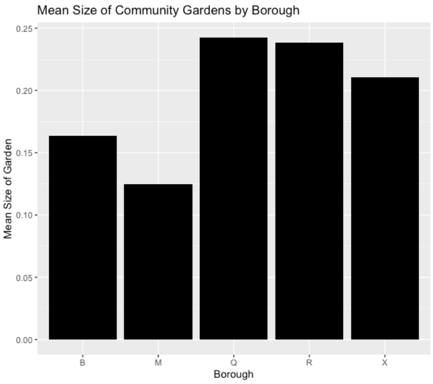
24 September 2019

As I live across the street from a community garden, I have often wondered about their operations—how does one become a member, what city department claims jurisdiction to these green spaces, and how many community gardens are there in New York City (NYC). For this descriptive analysis, I searched the NYC Open Data website for a data set on community gardens, and found the NYC Greentumb Community Gardens (NYCGCG) listing (Department of Parks and Recreation). This dataset, which was prepared by the Department of Parks and Recreation (DPR) contains 536 entries and seventeen variables—including Property ID (PropID), Borough (Boro), Community Board (Community.Board), Council District (Council.District), Garden Name (Garden.Name), Address, Size, Jurisdiction, Neighborhood Name (Neighborhood.Name), Cross Streets (Cross.Streets), Latitude, Longitude, Postcode, Census Tract (Census.Tract), Building Identification Number (BIN), Tax Lot (BBL), and Census Neighborhood Tabulation Area (NTA). And the data was last updated on 10 September 2018 (Department of Parks and Recreation).

At first, I thought with so many variables, the NYCGCG dataset would allow for a rich categorical analysis (Code in Appendix). However, as I explored the dataset, the most frustrating limitation was the fact that there are 1,099 NA values throughout the columns. This forced me to restrict my (already very narrow) analysis of the number and the size of the gardens throughout NYC. A further limitation to this dataset is the lack of documentation on how the data was sourced, the units of measure used for the size of the gardens, and coordinate system used to obtain the latitude and longitude.

Upon reflection, I should have selected a different dataset; but I thought that as there is a large amount of data in the world that is not perfect, the exercise of working with imperfect data would prove useful. In order to simplify matters, I eliminated variables that I could not reference without joining another dataset, and I excluded missing size observations. Most of the tables I created from the original dataset contained between two and ten variables; however, due to missing information, I found that only two variables provided a larger picture of the dataset—borough and size. I believed I could use additional variables, such as postcode and latitude and longitude to determine if there were any observations that were duplicates. However, this was not possible as inconsistent NA values made it difficult to determine which observation was correct.





Though incomplete and now curated data, I did observe that while 42.2% (226 gardens) of the 536 community gardens in NYC are located in Brooklyn, their average size is less than the average size of the community gardens located in Queens. It would be interesting to see if this finding correlates to any specific geographic or historical features of Queens and Brooklyn. Again, as 107 observations were eliminated from this study because they did not contain information on size, without correlating the NYCGCG dataset with other data, these limited findings require further review.

Appendix

```
#Pratt Info 640 Fall 2019
#Diedre Brown; dbrow207@pratt.edu
#Descriptive Data Analysis Project - Due 24 Sept 2019
#General Notes:
#NYC Greenthumb Community Gardens
(Source:https://data.cityofnewyork.us/Environment/NYC-Greenthumb-
Community-Gardens/ajxm-kzmj)
#The NYC Greenthumb Community Gardens (NYCGCG) dataset was obtained
from
NYC Opendata.
#The NYCGCG was last updated by the Department of Parks and Recreation
(DPR) on September 10, 2018.
#No information was provided as to why the data was collected, how the
data was collected, what each record represents, how this data can be
used, and any idiosyncrasies or limitations of the data that the user
should be made aware of.
#From NYC Parks/NYC Open Data:
#Established in 1978, NYC Parks GreenThumb is proud to be the nation's
largest urban gardening program, assisting over 550 gardens and over
20,000 volunteer gardeners throughout New York City.
#GreenThumb gardens create hubs of neighborhood pride and provide a
myriad of environmental, economic and social benefits to the
neighborhoods in which they thrive.
#GreenThumb's mission is to educate and support community gardens and
urban farming across the five boroughs, while preserving open space.
#By providing free garden materials, technical assistance, educational
workshops, and seasonal programs, GreenThumb supports neighborhood
```

volunteers who steward community gardens as active resources that strengthen communities.

```
#call libraries
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
#import csv dataset
nyccomgard <-read.csv("NYC_Greenthumb_Community_Gardens.csv")
#Impressions of data:
class(nyccomgard)
str(nyccomgard) #NYCGCG is categorical data. The set contains 536
observations; 17 variables; 1099 NA's.
summary(nyccomgard) #This showed that the Brooklyn (Boro B) has the
highest number of community gardens in NYC.
glimpse(nyccomgard)
head(nyccomgard)
tail(nyccomgard)
sum(is.na(nyccomgard)) #There are NA's everywhere. I am going to
exclude
the NA locations from any calculations because there is not enough
documentation to postulate on how to compensate for the missing values.
#PROBLEMS:
#Though size of the garden is listed, there is no documentation as to
what units this size is in (sf, acre, sm, etc.).
#Community Board values are a mix of alphanumeric information. As there
is a column for Borough and the Census.Tract and NTA infomation, the
alpha character of the Community Board values is redundant. I am not
sure how to break an alphanumeric character when there is no separater.
```

```
#Though Latitude and Longitude are given, there is no documentation
which states what coordinate system the data is referring to.
#PLAN:
#1. Manipulate data to remove variables not needed for this analysis
(PropID, Council.District, Garden.Name, Address, Cross.Streets,BIN,
BBL)
#2. Since there is a lot of missing information, remove NAs as much as
possible.
#3. Visualize/Compare Average size of community gardens by borough.
#Total number of community gardens in NYC, gardens per boro, and
percentages of gardens per boro
numNYCgard <- nyccomgard %>%
filter(!is.na(Boro)) %>%
group_by(Boro)%>%
count()
numNYCgard
totalNYCgard <- sum(numNYCgard$n)
totalNYCgard
perborogard <- numNYCgard %>%
mutate(perctbgard = ((n/536)*100) )
perborogard
ggplot(perborogard, aes(x=Boro, y=perctbgard)) +
geom_bar(stat = 'identity',
fill = "black",
size = 0.5) +
labs(x="Borough", y="Percentage of Gardens", title = "Percentage of
Gardens Per Borough in NYC (Total Gardens in NYC = 536)")
#Data Cleaning and Distributions
#Remove columns not needed for this analysis
```

```
nycgardshort <- nyccomgard%>%
select(-PropID, -Council.District, -Garden.Name, -Address, -
Cross.Streets, -BIN, -BBL)%>%
filter(Size != 'NA')%>%
arrange(Size)
summary(nycgardshort)
glimpse(nycgardshort)
head(nycgardshort)
#Histogram of Average Size of Community Gardens by Boro
borosizegrouped <- nycgardshort%>%
group_by(Boro)%>%
summarize(meanSize = mean(Size))
summary(borosizegrouped)
meanborogsize<-ggplot(borosizegrouped,aes(x=Boro, y=meanSize))+
geom_bar(stat = "identity",
fill = "black",
size = 0.5)+
labs(x="Borough", y="Mean Size of Garden", title = "Mean Size of
Community Gardens by Borough")
meanborogsize
#Community Gardens Jurisdiction Per Borough
nycgardjur <- nycgardshort %>%
filter(NeighborhoodName != "")%>%
group_by(Boro, Jurisdiction)
summary(nycgardjur)
glimpse(nycgardjur)
head(nycgardjur)
#I wanted to visualize this but realized after distributing that there
are more NAs and possible duplications, so I was unsure of what
```

information is correct.

References

Department of Parks and Recreation. (2017, September 16). NYC Greenthumb Community Gardens: NYC Open Data. Retrieved from https://data.cityofnewyork.us/Environment/NYC-Greenthumb-Community-Gardens/ajxm-kzmj