Hmk4-Q3

Yawen Han (yh3069)

11/14/2018

Question 3

Begin the analysis of one variable in the dataset you are using the final project. As this is an individual homework assignment, each group member should choose a different variable. Choose three visualizations as appropriate to show the distribution of the variable, conditioned on another variable if desired (for example, the distribution of income by region). Write a few sentences describing what you found and what new questions your visualizations have generated. (Faceted graphs count as one graph; graphs put together with grid.arrange() or similar count as multiple graphs.)

The dataset in my group's final project is "Street tree data" from the "TreesCount 2015 Street Tree Census", conducted by volunteers and staff organized by NYC Parks & Recreation and partner organizations. The link to the dataset is https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh (https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh).

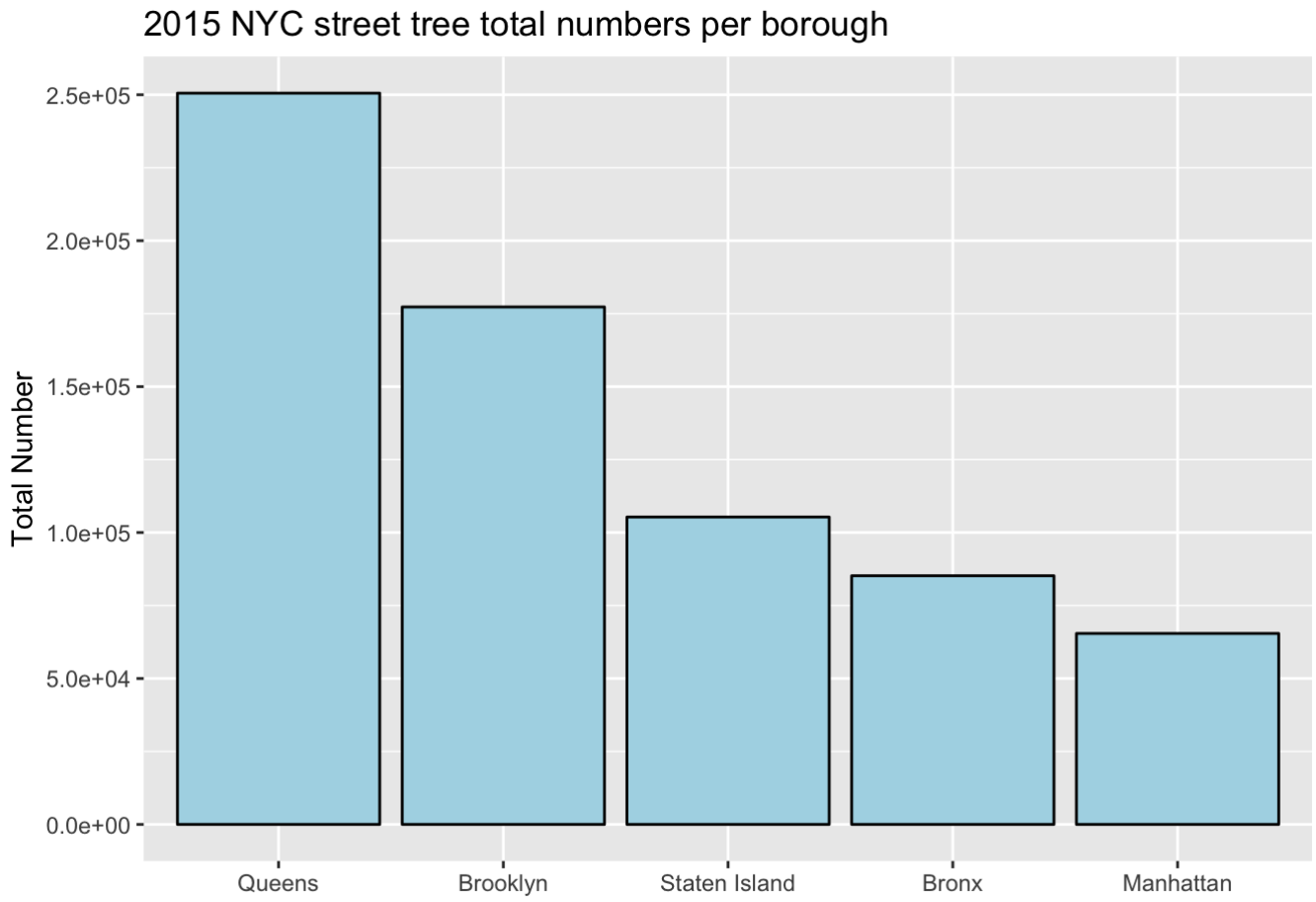
The variable I choose from the dataset is "borough", which gives the name of NYC borough in which tree point is located. It's a categorical variable with 5 class: Queens, Brooklyn, Staten Island, Manhattan, and the Bronx.

To explore the distribution of variable "borough", the following three visualizations are generated:

1. Bar Plot

Plot the barchart to show the number of street trees for each borough in NYC 2015. The plot is shown in decreasing order of total numbers.

```
library(tidyverse)
library(scales)
tree<-read.csv("/Users/yawenhan/Downloads/2015_Street_Tree_Census_-_Tree_Data.csv")
b<-tree %>% group_by(borough) %>% summarise(n=n())
# plot the barchart of total tree numbers by borough with a decreasing order
ggplot(b,aes(x=fct_reorder(borough,n,.desc=TRUE),y=n))+
  geom_bar(color="black",fill="lightblue",stat="identity")+
  ylab("Total Number")+scale_y_continuous(labels = scientific)+
  ggtitle("2015 NYC street tree total numbers per borough")+xlab("")
```



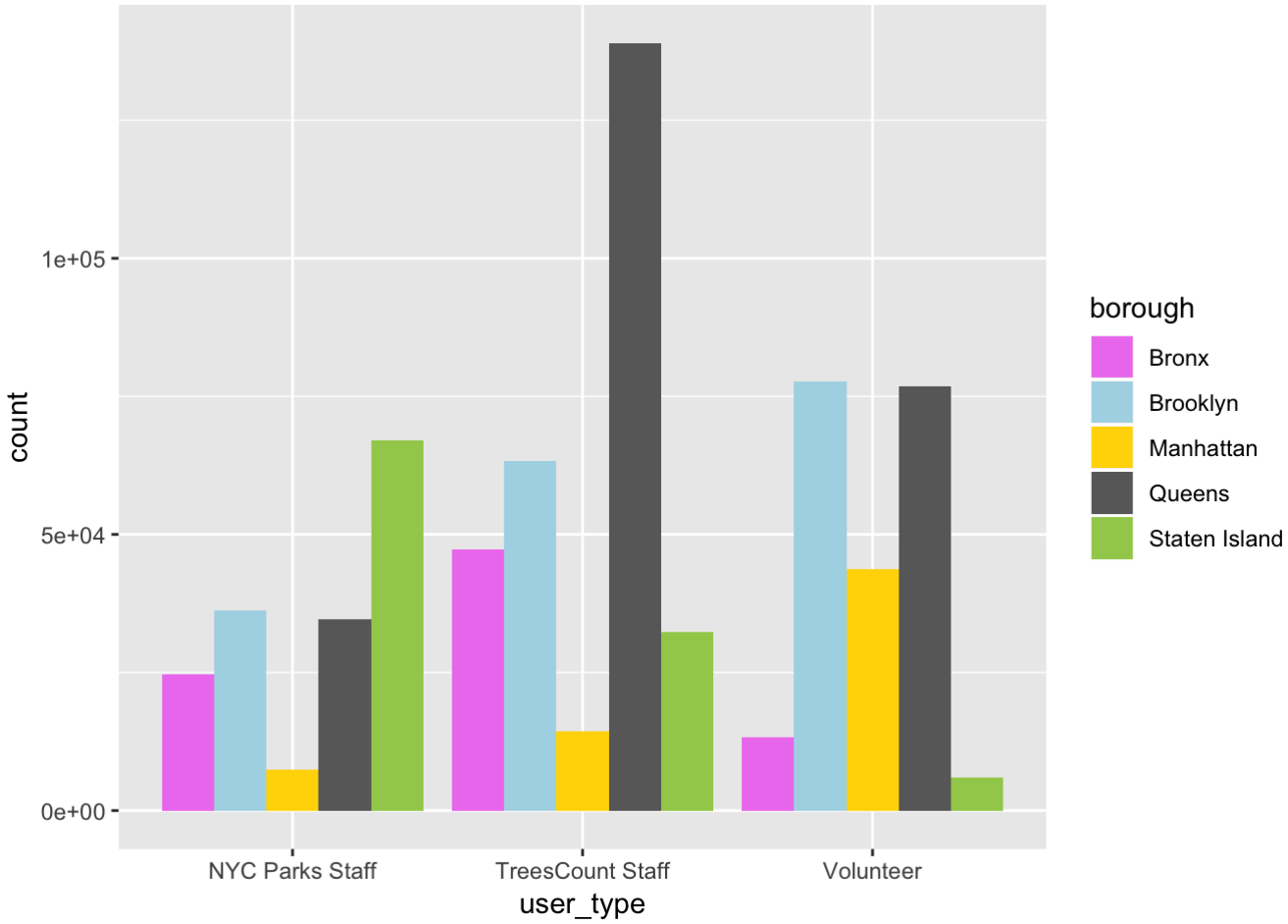
According to the bar chart above, the total number order of NYC borough is 'Queens', 'Brooklyn', 'Staten Island', 'Bronx' and 'Manhattan'. In year 2015, among all five boroughs, "Queens" had the largest number of street trees around 250,000, while "Manhattan" has the smallest number of street trees around 70,000.

More question: what influence the total number of street trees?

2.Grouped bar chart

Plot the grouped bar plot to explore the distribution of borough based on user_type. "User_type" describes the category of user who collected this tree point's data. The order of bough in each group is based on the order of total numbers from (1).

```
ggplot(tree, aes(x = user_type, fill = borough)) +
    geom_bar(position = "dodge") +
  scale_fill_manual(values = c("violet","lightblue","gold","dimgray","darkolivegreen3")
 )
```



From the grouped bar plot above, the distribution of borough based on user_type are compared as follows:

1)The distributions of borough street trees for the different user_type group is different;
2)The count of borough street trees are also different for different user-type;
3)“Queens” has most street trees and most of its user_type is “TreesCount Staff”, and “Manhatten” has the least street trees and most of its user_type is “Volunteer”.

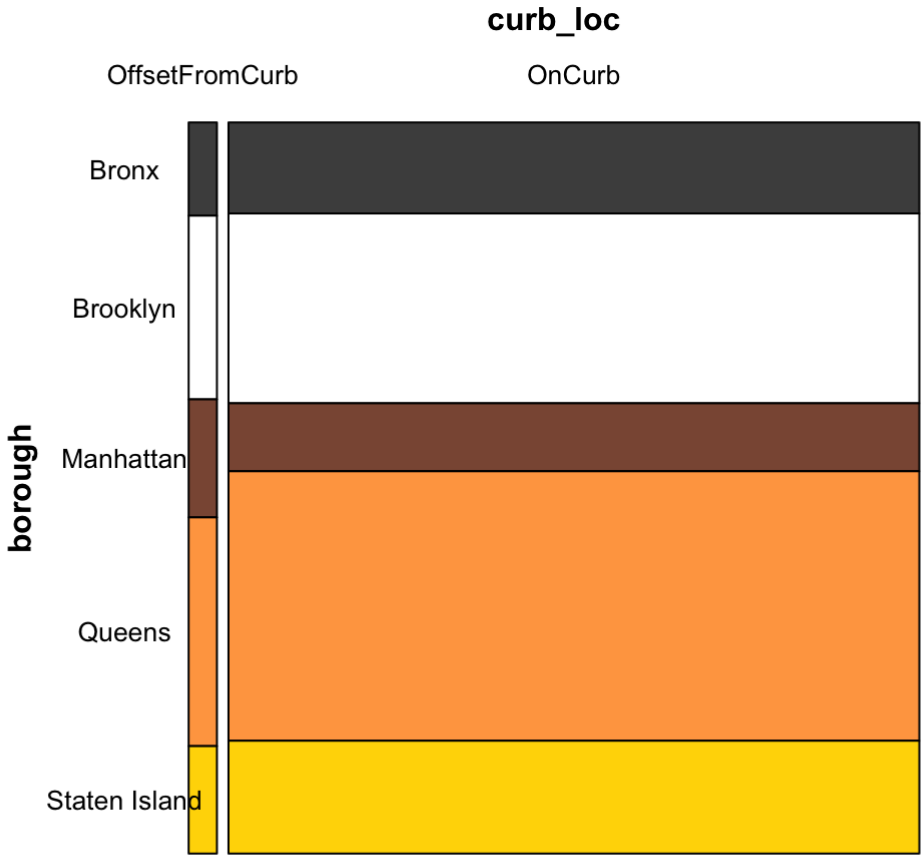
In conclusion, considering the above observations with the plot 1, The distributions of borough street trees do not follow the distribution of ungrouped borough street trees(shown in plot 1).

More question: is there any correlation between user_type and the borough count of street trees?

3.Mosaic plot

Plot the mosaic chart to explore the distribution of borough based on curb_loc. “Curb_loc” describes Location of tree bed in relationship to the curb; trees are either along the curb (OnCurb) or offset from the curb (OffsetFromCurb). The order of bough in each group is based on the order of total numbers from (1).

```
library(vcd)
library(grid) # needed for gpar
fillcolors <- c("gray30","white","lightsalmon4","tan1","gold")
vcd::mosaic( borough ~ curb_loc, tree, gp = gpar(fill = fillcolors), direction = c("v",
"h"),tl_labels = c(TRUE, TRUE),labeling = labeling_border(gp_labels = gpar(fontsize = 10
),gp_varnames = gpar(fontsize = 12,fontface = 2), rot_labels = c(0, 90, 0, 0), offset_va
rnames = c(0.7,0,0,2.4), offset_labels=c(0.5,0,0,1), pos_labels = c("center", "center",
"left", "center")))
```



From the mosaic plot above, the distribution of borough based on curb_loc are compared as follows:

1)The distributions of borough street trees for different curb_loc group do not have a significant difference;
2)The relative ratio of “Queens” street trees with “OffsetFromCurb” is smaller than that with “OnCurb”;
3)The relative ratio of “Manhatten” street trees with “OffsetFromCurb” is greater than that with “OnCurb”;

In conclusion, considering the above observations with the plot 1, although there is a little difference of distributions in between “OffsetFromCurb” and “OnCurb” group, the difference is not significant. Thus, no significant correlations between “borough” and “curb_loc”.

More question: what else variables might affect the distribution of the borough? for example: area, population.