NYC Trees Analysis

Bingquan Wu, Dazun Sun, Hangyu Zhou, Xiaoyu Chen

12/10/2018

Contents

Introduction **2**

Description of Data **2**

Analysis of Data Quality **2**

Main Analysis **4**

Number of trees 4
Breast height diameter of trees 9
Tree species 12
Health status 14
Links to codes 19

Executive summary **19**

Interactive component **22**

Conclusion **23**

Introduction

New York City, as one of the busiest city in the world, has surprisingly large amount of green space. This draws our attention to analyze the trees in NYC. Before looking at any data, the problems we are interested in include

- how many trees are there in NYC and how are they distributed?
- what are the most commonly planted species and why?
- what are the health conditions of the trees and what are influencing factors?

Our team members are their include:

- Bingquan Wu: contributes 8 graphs
- Dazun Sun: makes interactive component and 3 graphs
- Hangyu Zhou: analyzes data quality and finalizes report
- Xiaoyu Chen: cleans data and contributes 3 graphs

Description of Data

The data that we use is the citywide street tree data from the Street Tree Census in 1995, 2005 and 2015, conducted by volunteers organized by NYC Parks & Recreation. We access the datasets from the NYC OpenData website by direcly downloading them. Our main analysis is done using the tree data from 2015, the data from 1995 and 2005 are used to analyze the growth of number of trees. In addition, we use the tree cover rate data from wikipedia for major cities in US, the links for the data source is in this github file. Finally, in order to analyze the density of trees in each zipcode, we need the area of each zipcode. We use the ZIP Code Tabulation Area (ZCTA) dataset, and the data source is in this github file.

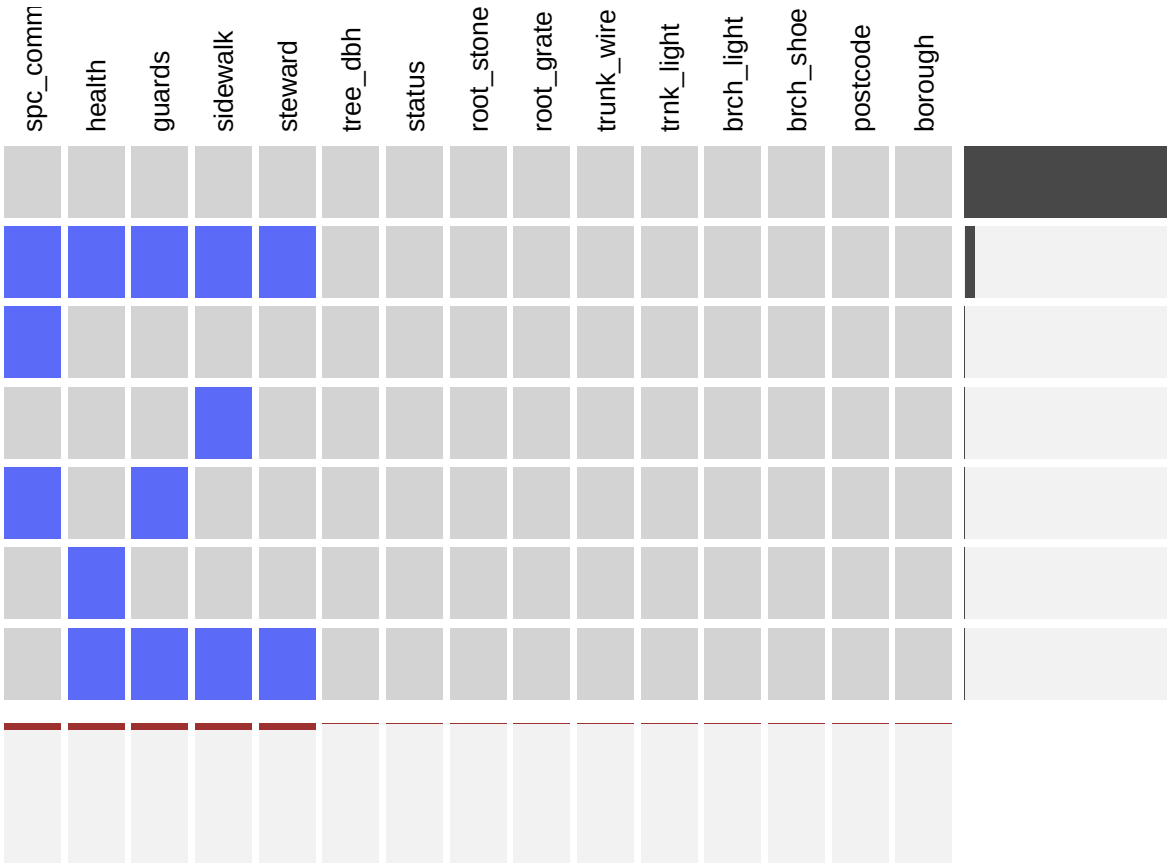
The 2015 tree data set has approximately $684,000$ rows, each representing a tree, and 45 columns, most of which are categorical. The only two meaningful numerical variables are the the diameter at breast level and the diameter of stump.

Analysis of Data Quality

To analyze the data quality, we first read our data from a local file and select only the variables we want to consider and then apply the visna function to extract the missing value paterns.

```
myvars <- c("tree_dbh", "status", "health", "spc_common",
            "steward", "guards", "sidewalk", "root_stone",
            "root_grate", "trunk_wire", "trnk_light",
```

```
            "brch_light", "brch_shoe", "postcode", "borough")
mydata <- mydata[myvars]
extracat::visna(mydata, sort = 'b')
```



Our data set has very few missing values and when looking at the original csv file, it seems that the missing values are due to the tree being dead or stump. To confirm our hypothesis, we make a bar plot of missing value by variable, grouped by status.

```
missing_value <- mydata %>% group_by(status) %>%
  summarize(spc_common = sum(is.na(`spc_common`)),
            health = sum(is.na(`health`)),
            guards = sum(is.na(`guards`)),
            sidewalk = sum(is.na(`sidewalk`)),
            steward = sum(is.na(`steward`))) %>%
  gather(key = variable, value = count, -status)

ggplot(missing_value, aes(status, count)) +
  geom_bar(aes(fill = variable), position = "dodge",
           stat="identity") +
  labs(x = 'status', y = 'number of m issing values',
       title = 'Number of missing values per status') +
  scale_fill_manual(values=cbPalette) + th
```

The graph above confirmed our hypothesis. We can now conclude that our data is very clean and has very few missing values.

Main Analysis

Our graphical analysis will focus on four categories, which involves

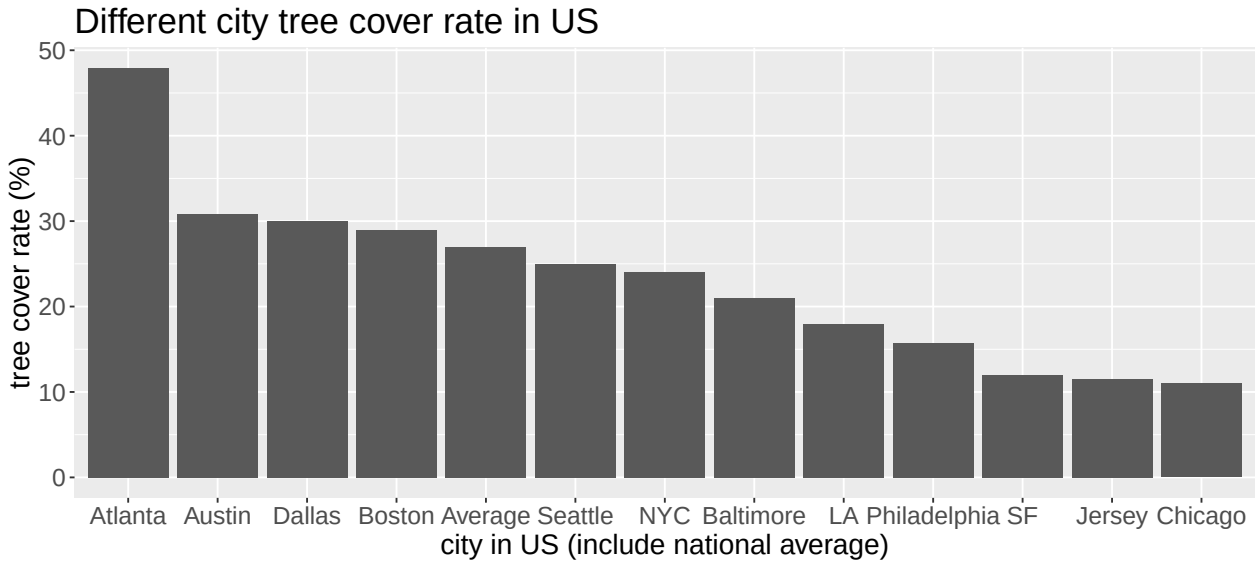
- number of trees
- breast height diameter of trees (measured at the height of an adult's breast, about 1.35m from ground)
- tree species
- health status

Number of trees

We will first compare the tree cover rate of NYC with other major cities in US.

```
ggplot(mytree,aes(x= reorder(Cities,-Tree.Cover),Tree.Cover)) +
  geom_bar(aes(y=Tree.Cover),stat='identity') +
```

```
  labs(title="Different city tree cover rate in US",
       x="city in US (include national average)",
       y = "tree cover rate (%)") +
  theme(plot.title = element_text(size=20)) +
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=16))
```

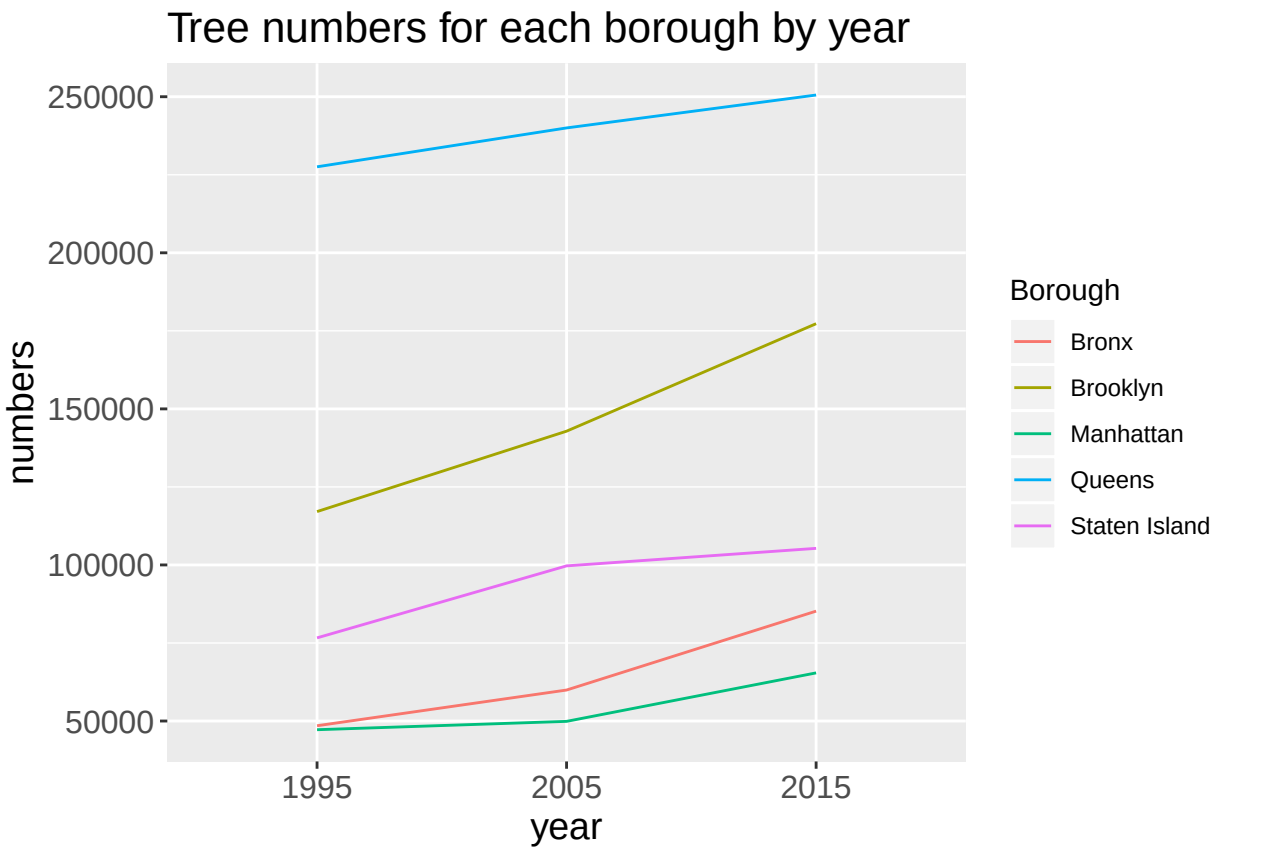


From the graph above, we can see that the national average of tree cover rate of cities in US is 27%, while New York City, is below the national average at 24%. So, we are trying to understand the trees in the New York City, the distribution and their changes through the time.

```
mydata <- mydata %>% mutate(health = replace_na(health, "Death"))
borough_2015 <- data.frame(table(mydata$borough))
colnames(borough_2015)[colnames(borough_2015) == 'Var1'] <- 'Borough'
borough_2015$Borough <- as.character(borough_2015$Borough)
borough_1995 <- data.frame(table(mydata_1995$Borough))
colnames(borough_1995)[colnames(borough_1995) == 'Var1'] <- 'Borough'
borough_1995$Borough <- as.character(borough_1995$Borough)
borough_2005 <- data.frame(table(mydata_2005$boroname))
colnames(borough_2005)[colnames(borough_2005) == 'Var1'] <- 'Borough'
borough_2005$Borough <- as.character(borough_2005$Borough)
borough_2005$Borough[borough_2005$Borough == "5"] <- "Staten Island"
borough_1995["Year"] <- "1995"
borough_2005["Year"] <- "2005"
borough_2015["Year"] <- "2015"
borough <- rbind(borough_1995, borough_2005, borough_2015)
colnames(borough)[colnames(borough) == 'Freq'] <- 'Numbers'

borough %>% ggplot(aes(x = Year,
```

```
                  y = Numbers,
                  group = Borough,
                  color = Borough)) +
  geom_line() +
  labs(title="Tree numbers for each borough by year",
    x = "year",
    y = "numbers") + th
```

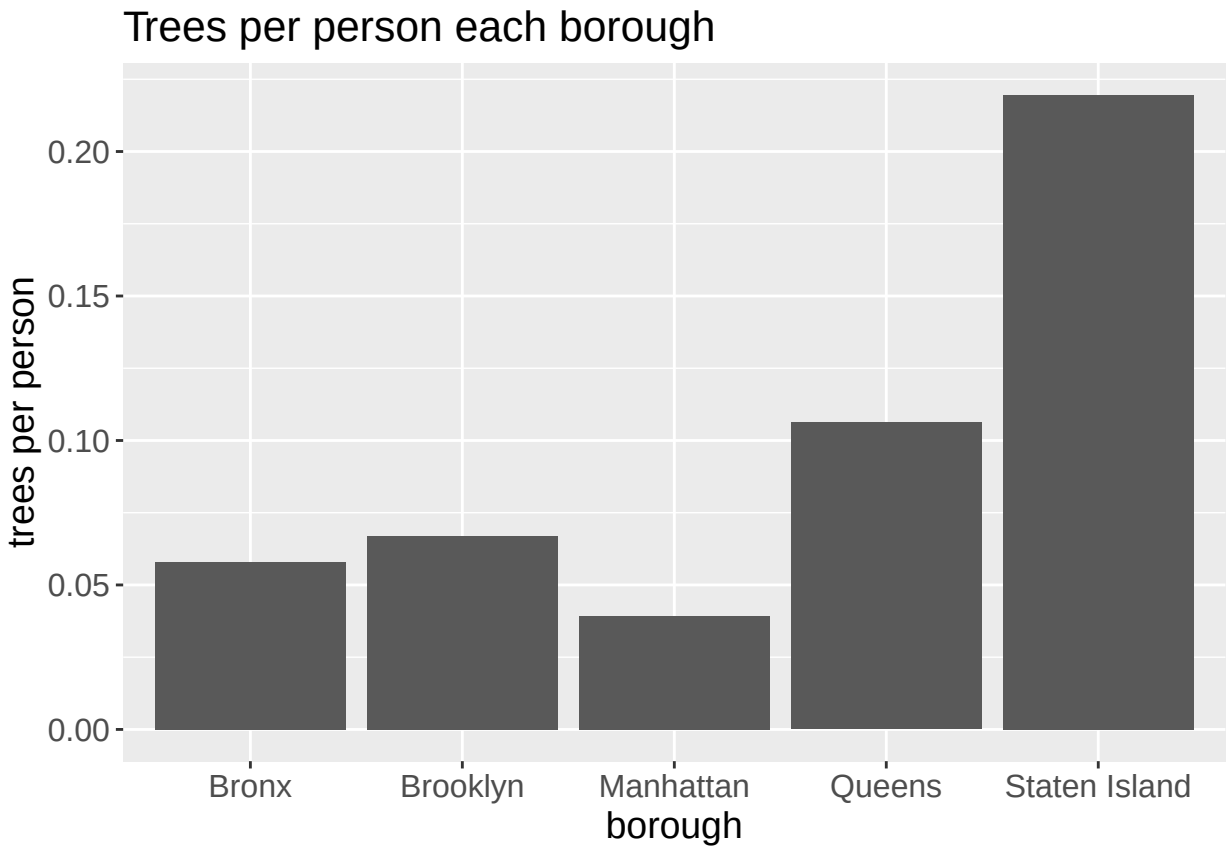


For five boroughs in New York City, number of trees varying by year intuitively reflect the green rate in each borough. It is evident to see from graph that number of trees show a trend of increasing. For the absolute tree numbers, Queens ranks first in five boroughs, followed by Brooklyn, Staten Island, Bronx and Manhatten. While for the increasing rate, Brooklyn ranks first, followed by Bronx, Manhattan, Queens and Staten Island. However, since each borough has different population, we also need to consider the population in each borough to make the comparison more meaningful. We use the most recent(2015) data in the next graph.

```
borough<-c("Manhattan","Bronx","Staten Island","Brooklyn","Queens")
borough<-factor(borough)
tree<-c(65423,85203,105318,177293,250551)
population<-c(1664727,1471160,479458,2648771,2358582)
area<-c(22.83,42.1,58.37,70.82,108.53)
populationdensity<-population/area
```

```
treedensity<-tree/area
mydata2<-data.frame(borough, treedensity,
                    populationdensity, tree,
                    population)

ggplot(mydata2,aes(x=borough)) +
  geom_bar(aes(y=tree/population),stat="identity") +
  labs(title="Trees per person each borough",
       x="borough",
       y = "trees per person") + th
```

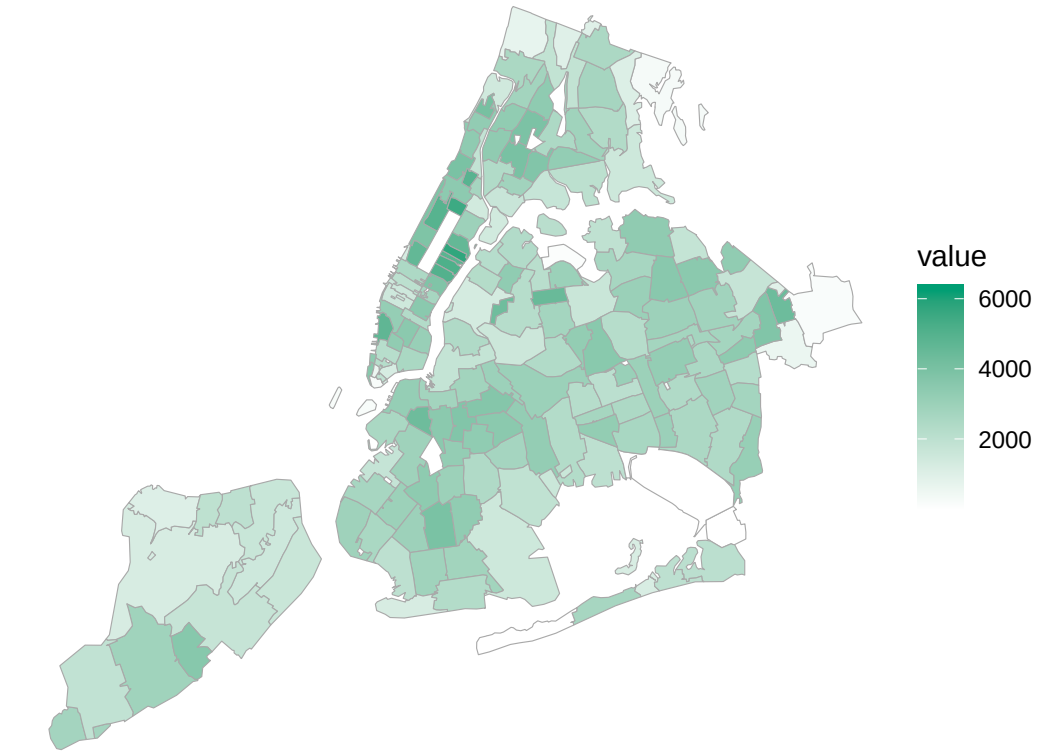


So, here is the plot of trees per person in each borough. Manhattan is still the lowesr, while Staten Island has the highest tree per person, much higher than other borough. This is mainly because, Staten Island has a low population density relatively, while Manhattan has a high population density. And also, Manhattan has much more buildings than other borough, and less place for planting trees.

In addition to trees per person, we are also interested in the density of trees in a fine-grained area. So we use the spatial heat map to visualize the density of trees in each zip code in 2015.

```
data("zip.regions")
valid_region <- zip.regions$region
valid_area <- tree_with_area %>%
```

```
  filter(land_area != 100)
density_tree <- valid_area %>%
  mutate(p_tree = 1.0/land_area) %>%
  group_by(postcode) %>%
  summarise(sum = sum(p_tree)) %>%
  mutate(region = as.character(postcode) , value = sum) %>%
  select(region, value) %>%
  filter(region %in% valid_region)
nyc_fips <- density_tree$region
zip_choropleth(density_tree,
               zip_zoom = nyc_fips,
               num_colors = 1,
               title = "Spatial heat map of the density of trees",
               legend = "Tree per sqmi") +
  scale_fill_gradient(low = "white", high = "#009E73")
```

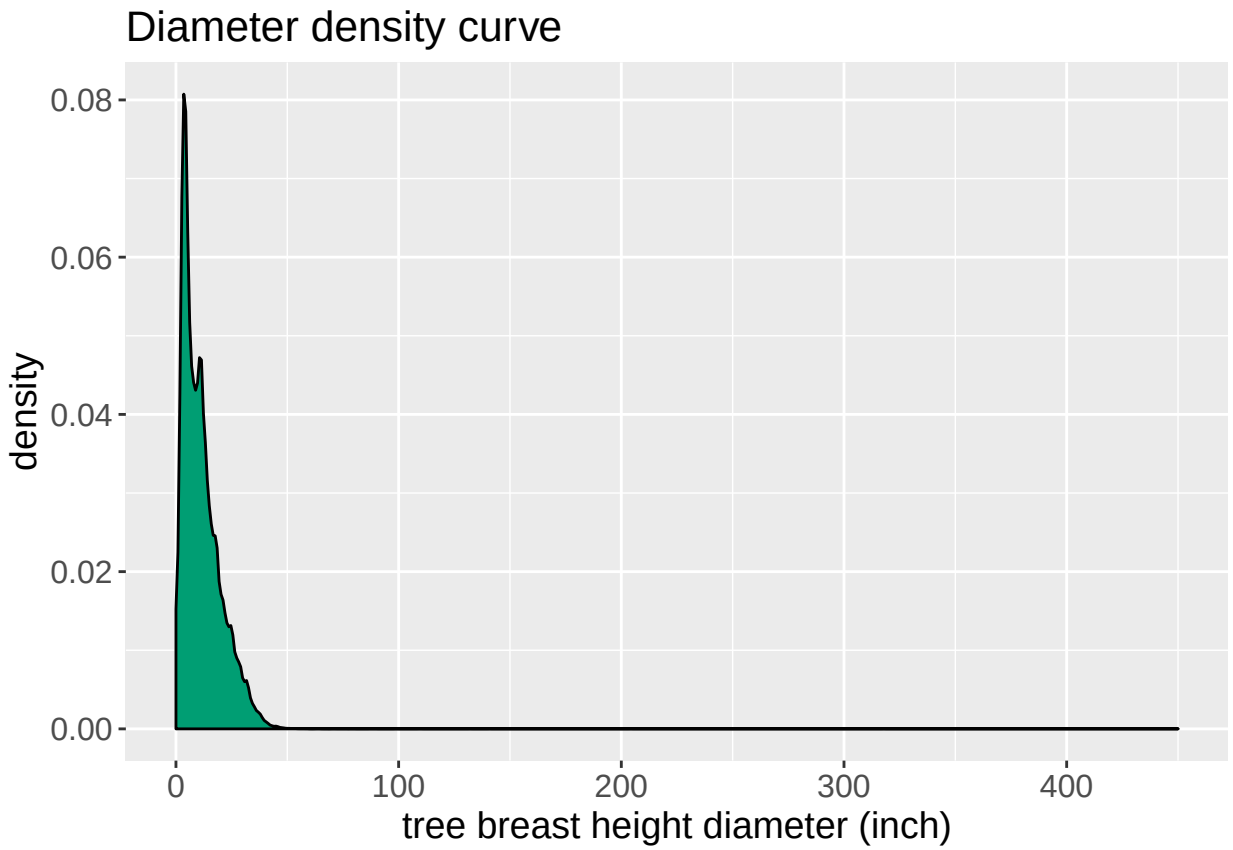


From the graph above, we can see that the area around central park have a quite high density of trees while in Midtown the color is much lighter. This may because there are more office buildings in Midtown. Besides, compared to the trees per person graph, we can find out that Manhattan's rank in tree density is better than that in tree per person, which validates the assumption that Manhattan has the highest population density.

Breast height diameter of trees

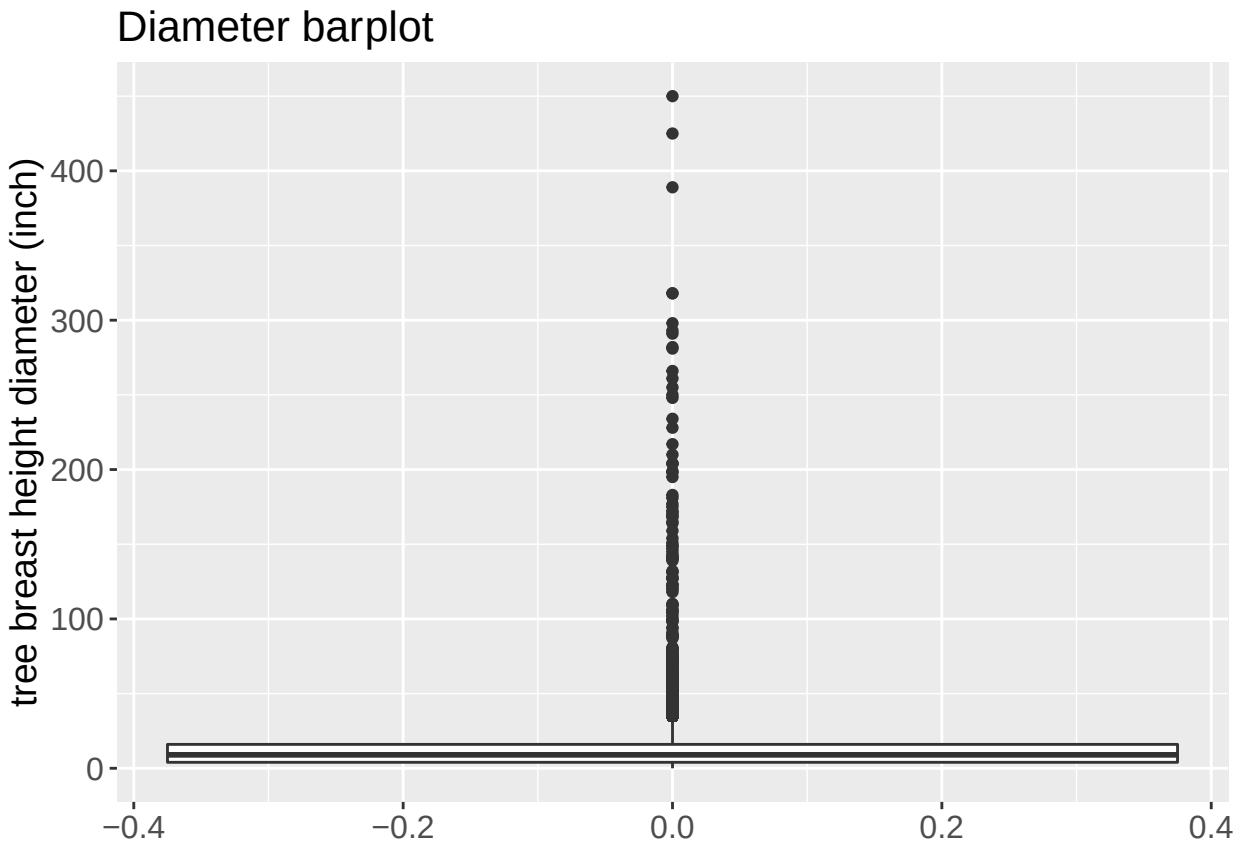
The dbh (breast height diameter) of trees is measured at the height of an adult's breast. It is often used to estimate the volume, biomass, and carbon storage of trees. The distribution of dbh of trees are as follow.

```
ggplot(mydata, aes(x=tree_dbh)) +
  geom_density(fill="#009E73") +
  labs(title="Diameter density curve",
       x="tree breast height diameter (inch)",
       y = "density") + th
```



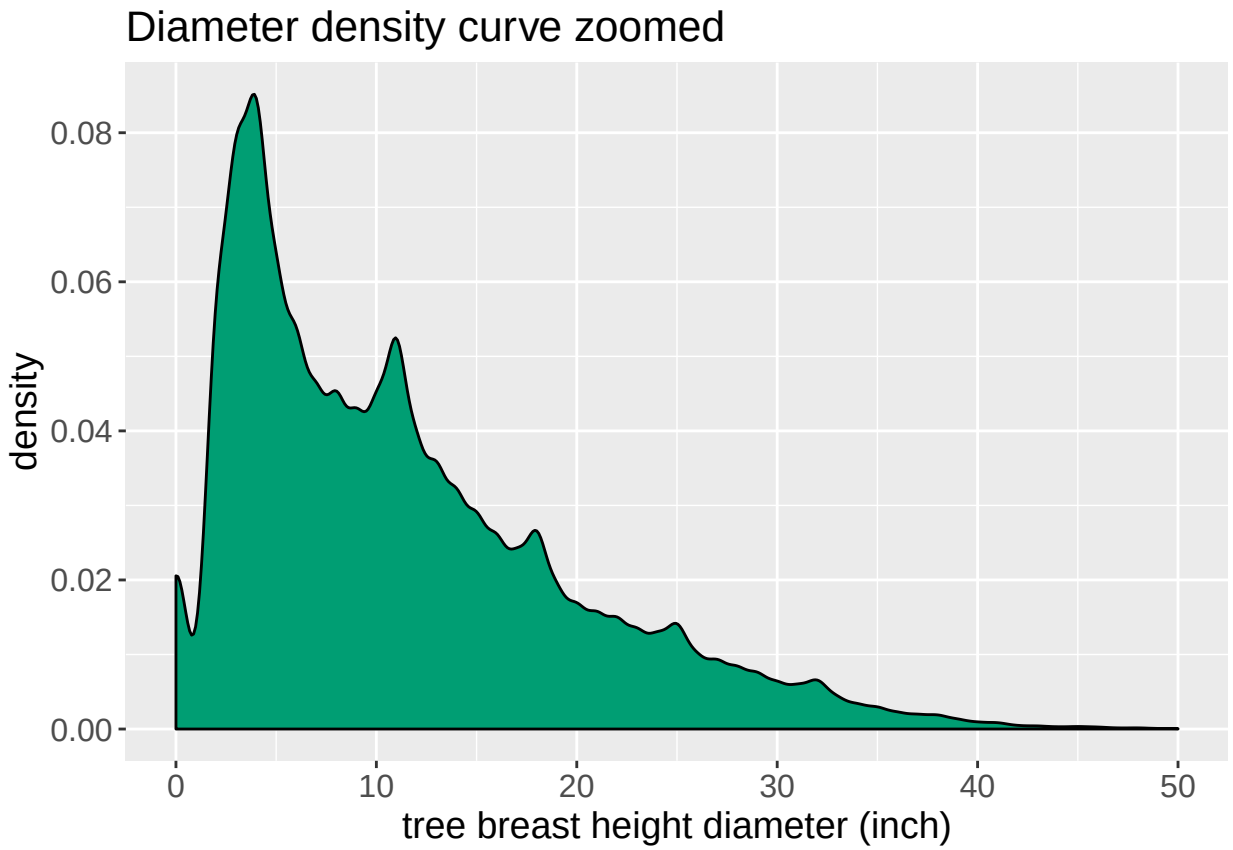
From the density curve we can see that most tree have a dbh from 0 to 30, and lots of them are in the range of 10-20. From the scale of x-axis, we can tell that there is few tree that has diameter of more than 400 which does not make sense. So the question now is, how do we decide whether a tree is considered as an outlier. The boxplot is useful in answering the question.

```
ggplot(mydata, aes(y=tree_dbh)) + geom_boxplot() +
  labs(title="Diameter barplot",
       y = "tree breast height diameter (inch)") + th
```



By plotting a boxplot. We can find out the outlier easily. We can see no trees has negative diameter, which makes sense. We can see that median is around 10, while 75% of data are from 0-20, and data more than 35 can be considered to be outliers. From now on, we will zoom in and focus on trees with bdh smaller than or equal to 50.

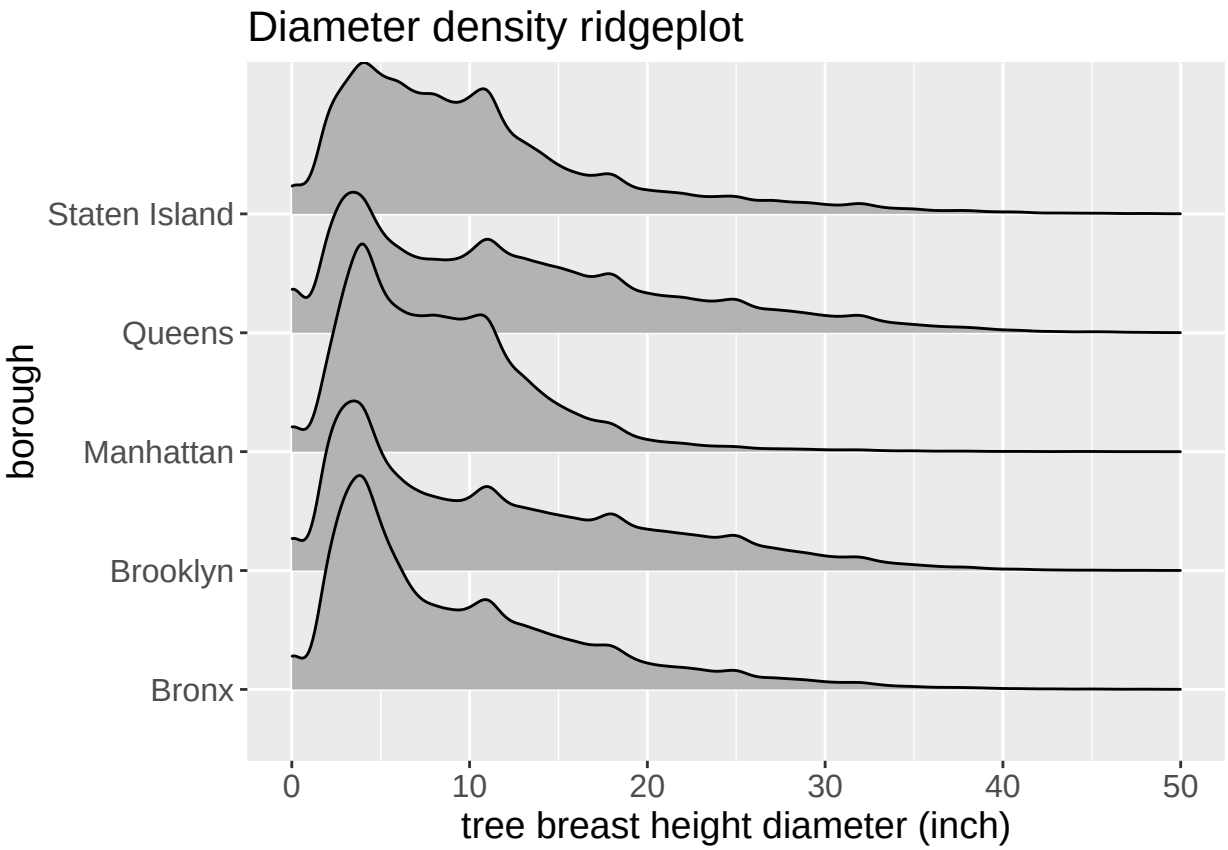
```
ggplot(mydata, aes(x=tree_dbh)) + geom_density(fill="#009E73") +
  labs(title="Diameter density curve zoomed",
       x="tree breast height diameter (inch)", y = "density") +
  xlim(0, 50) + th
```



Now, we get aside the outliers, and plot the density curve again. This time we will get a somehow zoomed view of the curve. We can clearly see the serval peaks. The highest peak is around 4cm, with a density of more than 0.08. In addition, we can see the diameter is distributed normally around the peak of 4cm, even though the plot does not look like a bell, since the lower bound of the diameter is 0 and do not have a upper bound.

Using a ridgeplot, we can compare the density distribution by borough very easily. Ridgeplot give us a direct understanding of different distribution, and easy to compare.

```
ggplot(mydata, aes(x = tree_dbh, y = borough)) + geom_density_ridges() +
  labs(title="Diameter density ridgeplot",
       x="tree breast height diameter (inch)",
       y = "borough") +
  xlim(0, 50) + th
```



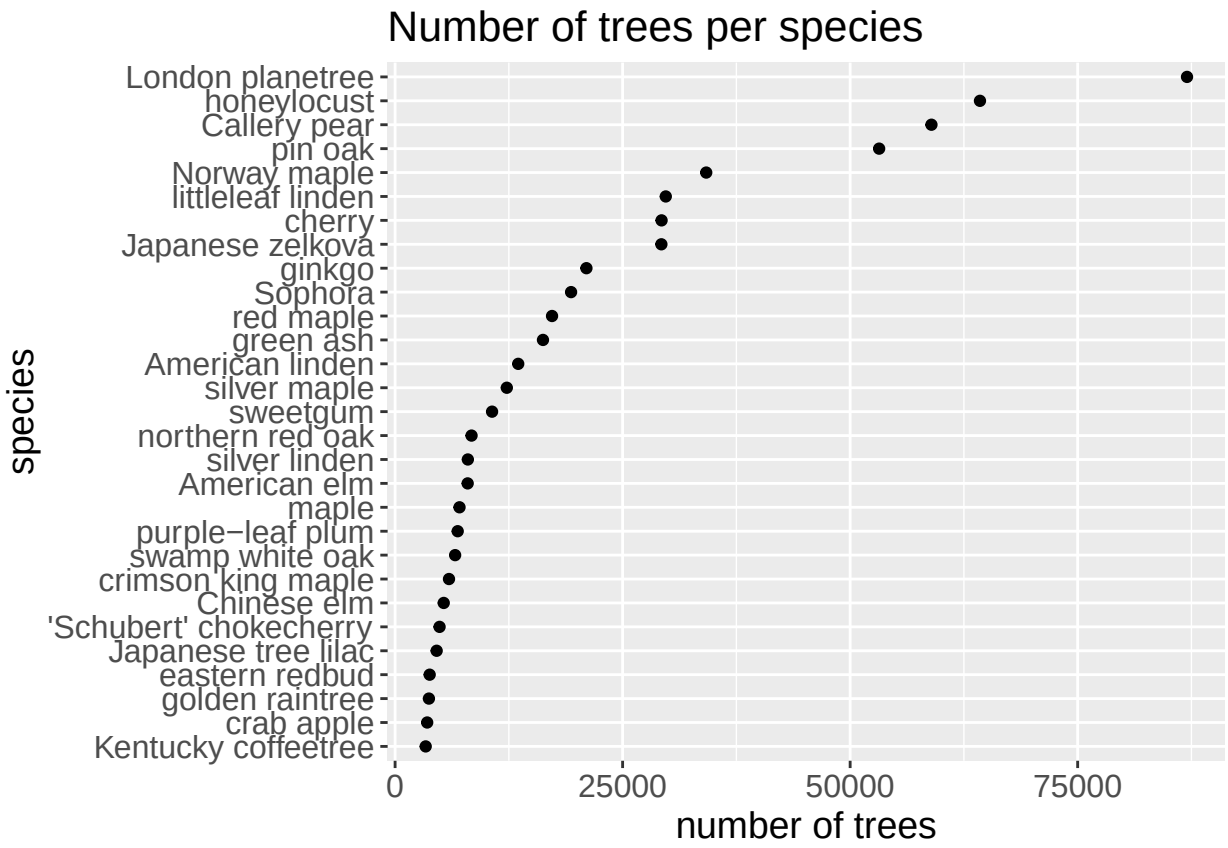
We can get many information, for example, Queens has more trees with diameter from 25-50 than Manhattan, which have very few tree with diameter in that range.

Also, we can find out a single-peak pattern in every borough, and a similar normal distributin of the pattern, which is useful in our further study if calculation will be involved.

Tree species

We will first look at the number of trees for each species in NYC.

```
spc <- mydata %>% group_by(spc_common) %>% summarise(count = n())
spc <- spc %>% top_n(30)
spc <- spc[-30,]
ggplot(spc, aes(x = count, y = fct_reorder(spc_common, count))) +
  geom_point() +
  labs(title="Number of trees per species",
       x = "number of trees",
       y = "species") + th
```

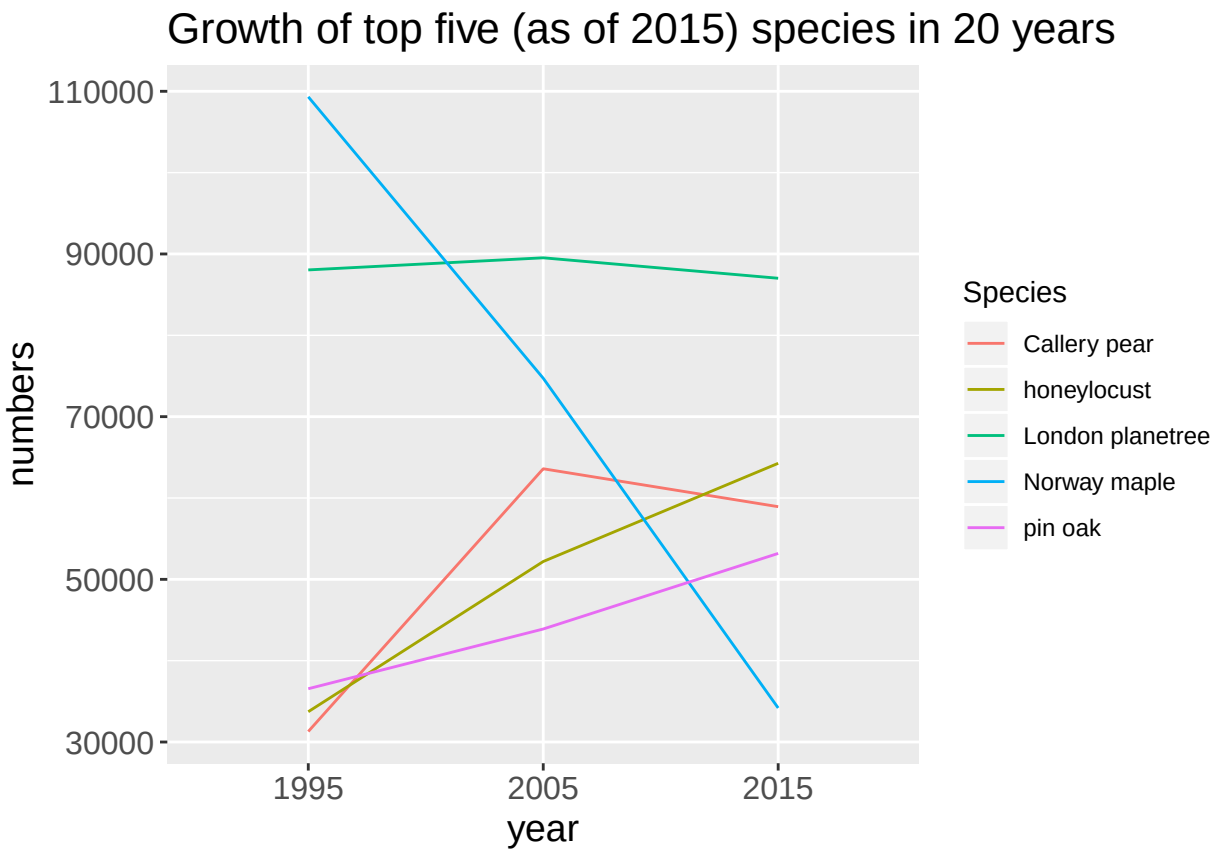


There are five species which are clearly planted more in NYC, namely, London planetree, honeylocust, callery pear, pink oak and Norway maple. It is interesting to find our how do the number of these trees change overtime.

```
sort(table(mydata$spc_common), decreasing=TRUE)
name2015 <- c('London planetree', 'honeylocust', 'Callery pear',
              'pin oak', 'Norway maple')
freq2015 <- c(87014, 64264, 58931, 53185, 34189)
spc2015 <- data.frame(name2015, freq2015)
colnames(spc2015)[colnames(spc2015) == 'name2015'] <- 'Species'
colnames(spc2015)[colnames(spc2015) == 'freq2015'] <- 'Numbers'
spc2015["Year"] <- "2015"
sort(table(mydata_1995$Spc_Common), decreasing=TRUE)
freq1995 <- c(88040, 33727, 31293, 36553, 109321)
spc1995 <- data.frame(name2015, freq1995)
colnames(spc1995)[colnames(spc1995) == 'name2015'] <- 'Species'
colnames(spc1995)[colnames(spc1995) == 'freq1995'] <- 'Numbers'
spc1995["Year"] <- "1995"
sort(table(mydata_2005$spc_common), decreasing=TRUE)
freq2005 <- c(89529, 52191, 63593, 43895, 74721)
spc2005 <- data.frame(name2015, freq2005)
colnames(spc2005)[colnames(spc2005) == 'name2015'] <- 'Species'
colnames(spc2005)[colnames(spc2005) == 'freq2005'] <- 'Numbers'
```

```
spc2005["Year"] <- "2005"
spc <- rbind(spc1995, spc2005, spc2015)
spc$Species <- as.character(spc$Species)
spc$Numbers <- as.integer(spc$Numbers)
```

```
spc %>% ggplot(aes(x = Year,
                   y = Numbers, group = Species, color = Species)) +
          geom_line() +
  labs(title="Growth of top five (as of 2015) species in 20 years",
              x = "year",
              y = "numbers") + th
```



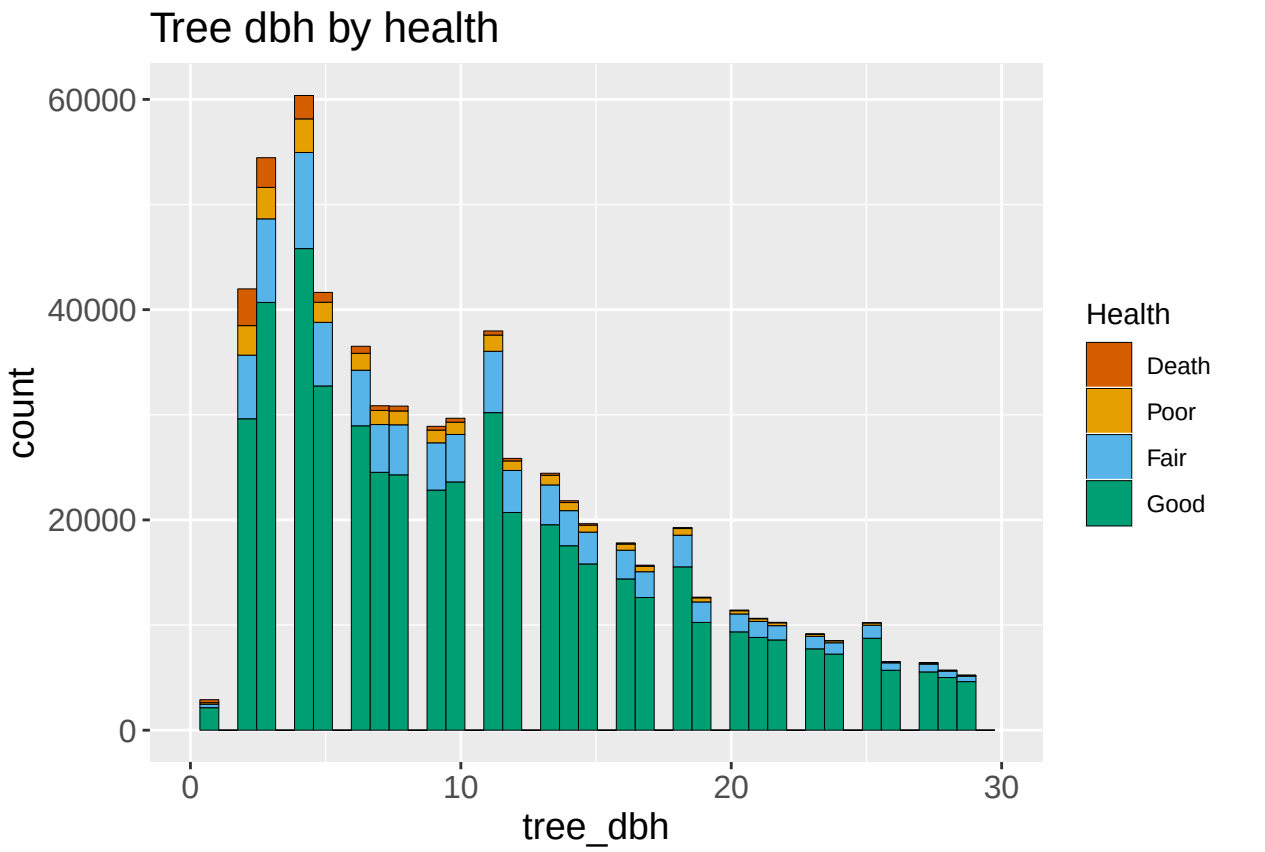
This graph is interesting and we will discuss it in the excutive summary section.

Health status

For the health status, our first hypothesis is that tree dbh is related to the health status.

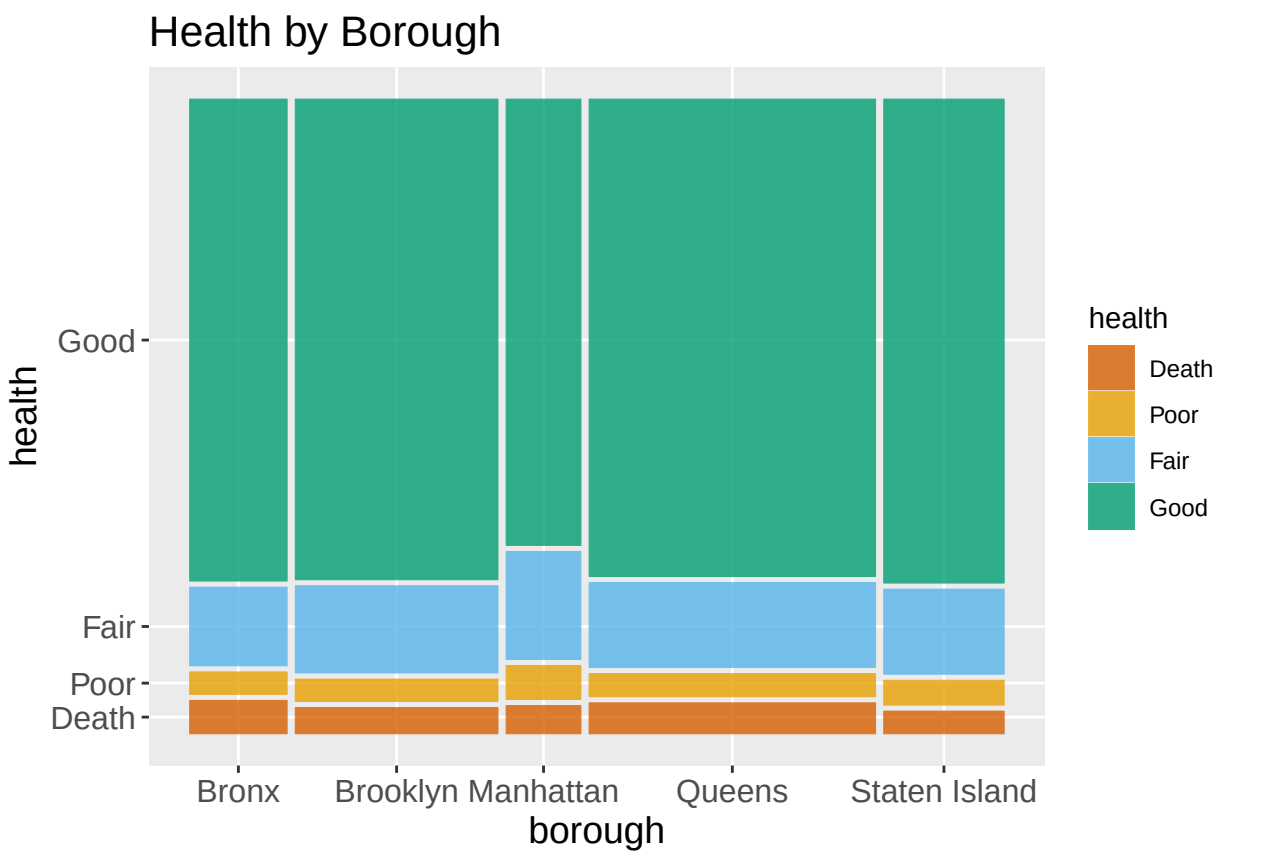
```
g <- ggplot(mydata, aes(tree_dbh)) + scale_fill_manual(values=cbPalette)
g + geom_histogram(aes(fill=factor(health, levels = c ("Death", "Poor",
                                                       "Fair", "Good"))),
                   binwidth = .7,
```

```
                  col="black",
                  size=.1) +  xlim(0,30) + ggtitle("Tree dbh by health") +
  labs(fill="Health") + th
```



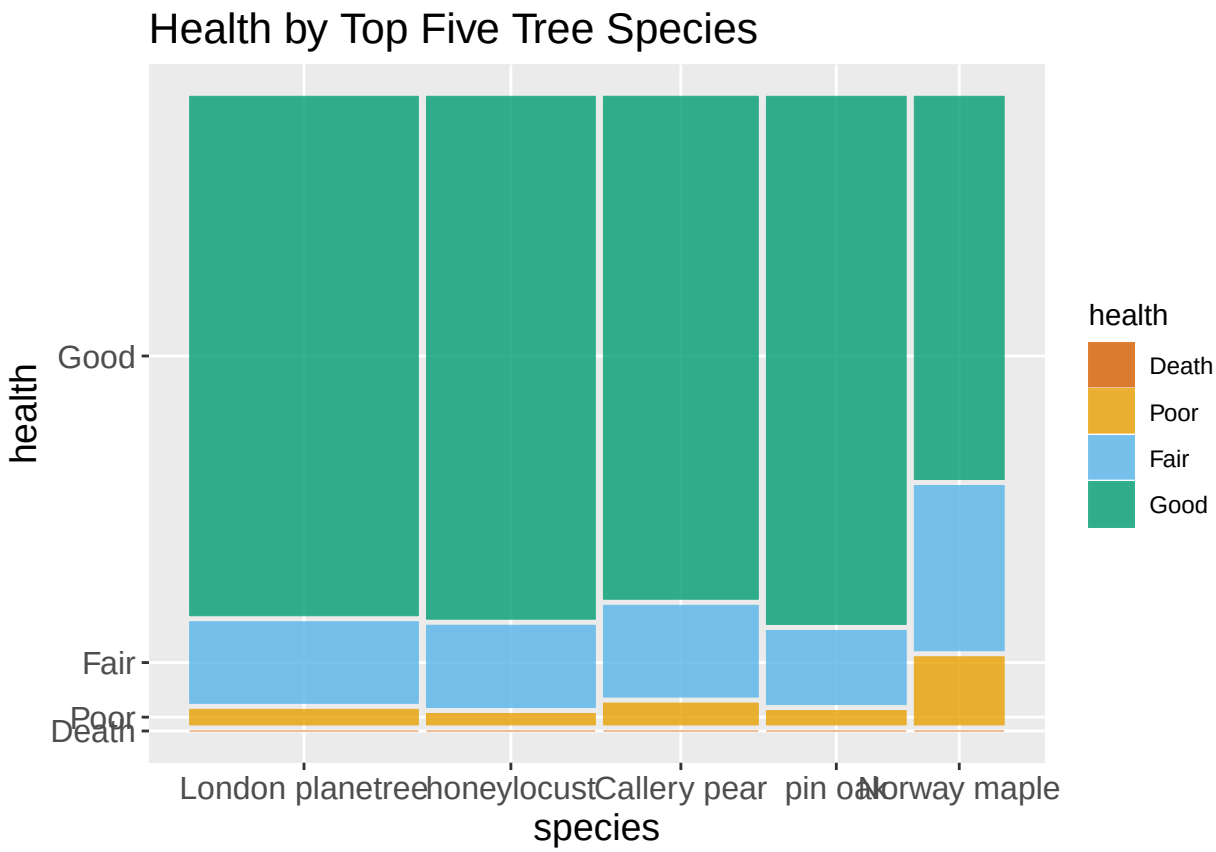
The graph descirbes the health status of each diameter of the trees in the dataset. It is evident that most trees are in good status and obviously, dead trees account only for a small fraction. From the plot, we also know that most trees fall in diameters below 20 inches and the status for each tree is ranked by Good, Fair, Poor and Death. Next, we will analyze the relationship between some factors and health.

```
df_tree_health <- mydata[!is.na(mydata$health), ]
df_tree_health$health <- factor(df_tree_health$health,
                                levels = c("Death", "Poor", "Fair", "Good"),
                                ordered = T)
ggplot(df_tree_health) +
  geom_mosaic(
    aes(x=product(health, borough),
        fill = health
        ),
    divider=c("vspine" , "hspine")
  ) + labs(x="borough", y="health") +
  ggtitle("Health by Borough") +
  scale_fill_manual(values=cbPalette) + th
```



From the above graph, we can find out that the propotion of trees in good health is visibly lower compared to other four boroughs. This is reasonable since the living environment is less friendly to trees in busy city.

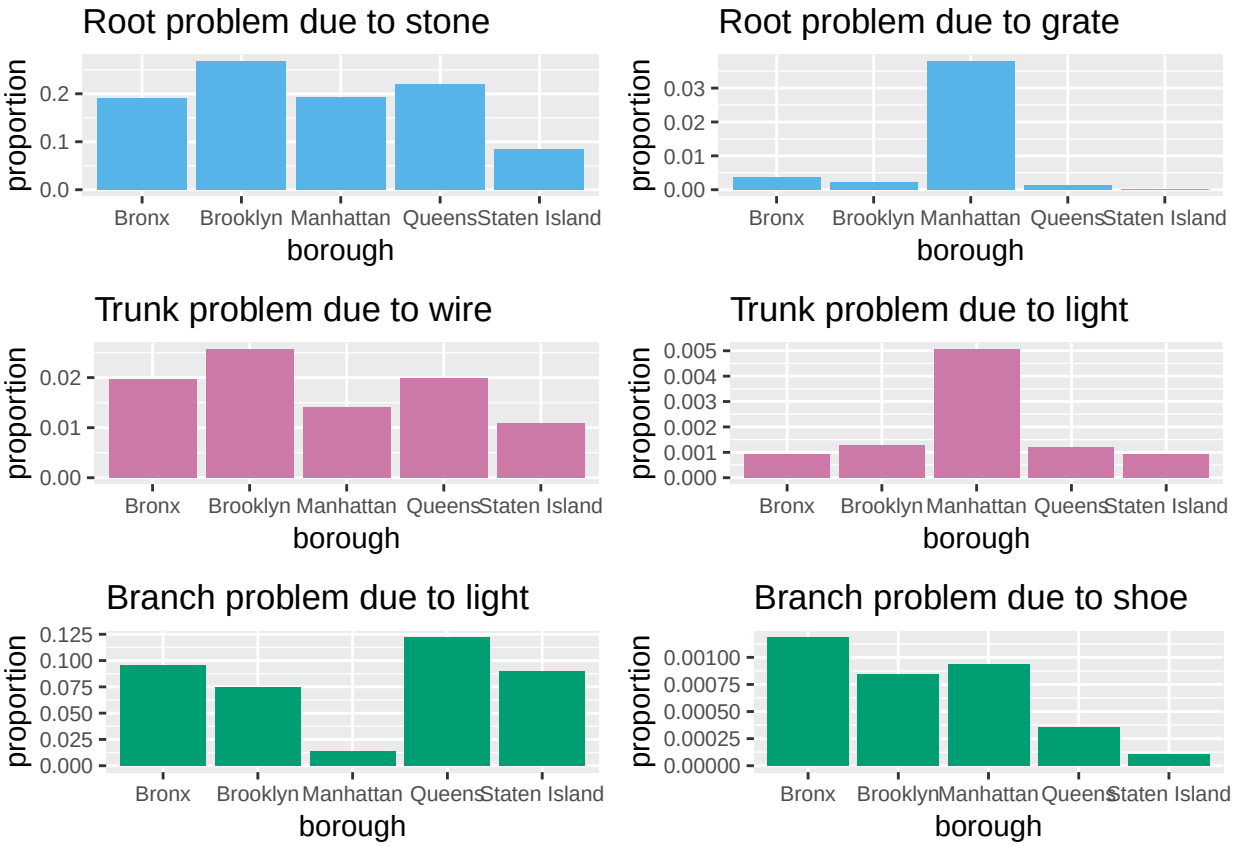
```
species_sum <- df_tree_health %>% group_by(spc_common) %>%
  summarise(s_sum = n()) %>%
  arrange(-s_sum)
top_five <- as.character(species_sum$spc_common[1:5])
df_tree_top <- df_tree_health %>%
  mutate(spc_common = ifelse(spc_common %in% top_five, spc_common, "Other")) %>%
  filter(spc_common != "Other")
df_tree_top$spc_common <- factor(df_tree_top$spc_common, levels = top_five)
ggplot(df_tree_top) +
  geom_mosaic(
    aes(x=product(health, spc_common),
        fill = health
        ),
    divider=c("vspine" , "hspine")
  ) + labs(x="species", y="health") +
  ggtitle("Health by Top Five Tree Species") +
  scale_fill_manual(values=cbPalette) + th
```



Different species of trees vary in durability. Since there are hundreds of species, we choose the five species with most amount. From the above graph we can see that Norway maple has the poorest health, followed by Callery pear, and pin oak is the most sturdy one. Then we will analyze different health problems for trees.

```
boroughrootstone<-mydata %>% group_by(borough) %>% count(root_stone) %>%
  mutate(proportion = n / sum(n))
boroughrootstone2<- subset(boroughrootstone,
                           boroughrootstone$root_stone == "Yes")
boroughrootgrate<-mydata %>% group_by(borough) %>% count(root_grate) %>%
  mutate(proportion = n / sum(n))
boroughrootgrate2<- subset(boroughrootgrate,
                           boroughrootgrate$root_grate == "Yes")
boroughtrunkwire<-mydata %>% group_by(borough) %>% count(trunk_wire) %>%
  mutate(proportion = n / sum(n))
boroughtrunkwire2<- subset(boroughtrunkwire,
                           boroughtrunkwire$trunk_wire == "Yes")
boroughtrunklight<-mydata %>% group_by(borough) %>% count(trnk_light) %>%
  mutate(proportion = n / sum(n))
boroughtrunklight2<- subset(boroughtrunklight,
                            boroughtrunklight$trnk_light == "Yes")
boroughbrchlight<-mydata %>% group_by(borough) %>% count(brch_light) %>%
  mutate(proportion = n / sum(n))
```

```
boroughbrchlight2<- subset(boroughbrchlight,
                           boroughbrchlight$brch_light == "Yes")
boroughbrchshoe<-mydata %>% group_by(borough) %>% count(brch_shoe) %>%
  mutate(proportion = n / sum(n))
boroughbrchshoe2<- subset(boroughbrchshoe,
                          boroughbrchshoe$brch_shoe == "Yes")
s1 <- ggplot(boroughrootstone2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#56B4E9")+
  labs(title="Root problem due to stone",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
s2 <- ggplot(boroughrootgrate2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#56B4E9")+
  labs(title="Root problem due to grate",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
s3 <- ggplot(boroughtrunkwire2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#CC79A7")+
  labs(title="Trunk problem due to wire",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
s4 <- ggplot(boroughtrunklight2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#CC79A7")+
  labs(title="Trunk problem due to light",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
s5 <- ggplot(boroughbrchlight2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#009E73")+
  labs(title="Branch problem due to light",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
s6 <- ggplot(boroughbrchshoe2,aes(x=borough))+
  geom_bar(aes(y=proportion),stat="identity",fill="#009E73")+
  labs(title="Branch problem due to shoe",
       x="borough", y = "proportion") +
  theme(axis.text=element_text(size=8))
grid.arrange(s1, s2, s3, s4, s5, s6, ncol=2)
```



This graph is very informative and we will talk about it in the next section.

Links to codes

The github repo can be found here. The useful codes are in the final rmb file which can be found here. The codes for the interactive part can be found here

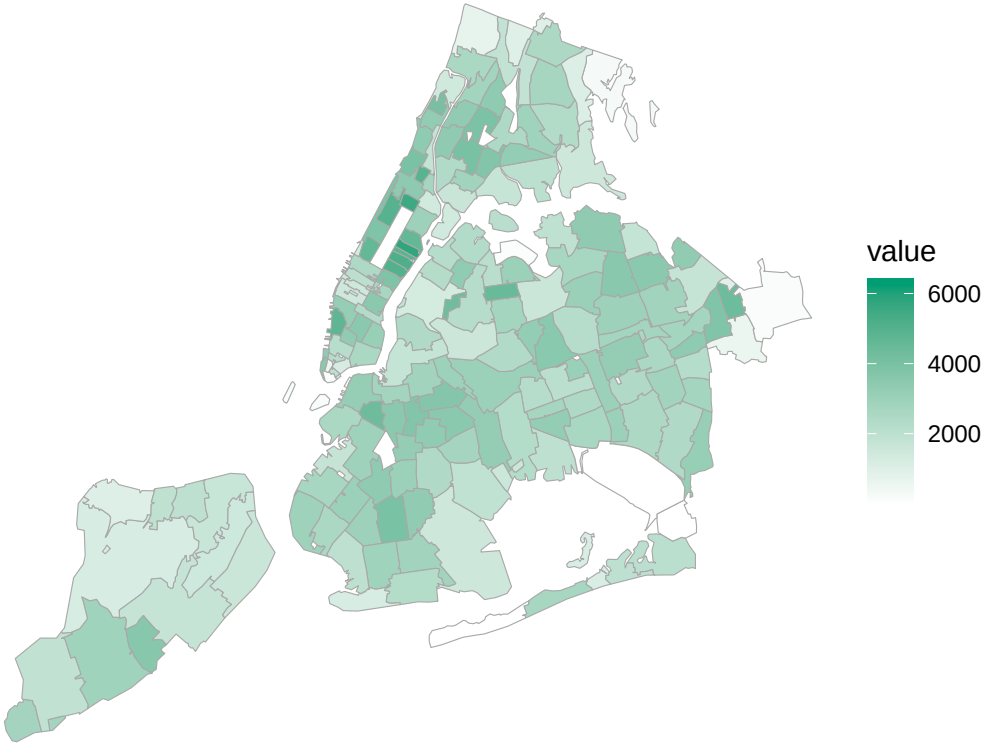
Executive summary

The three graphs that we find most revealing are

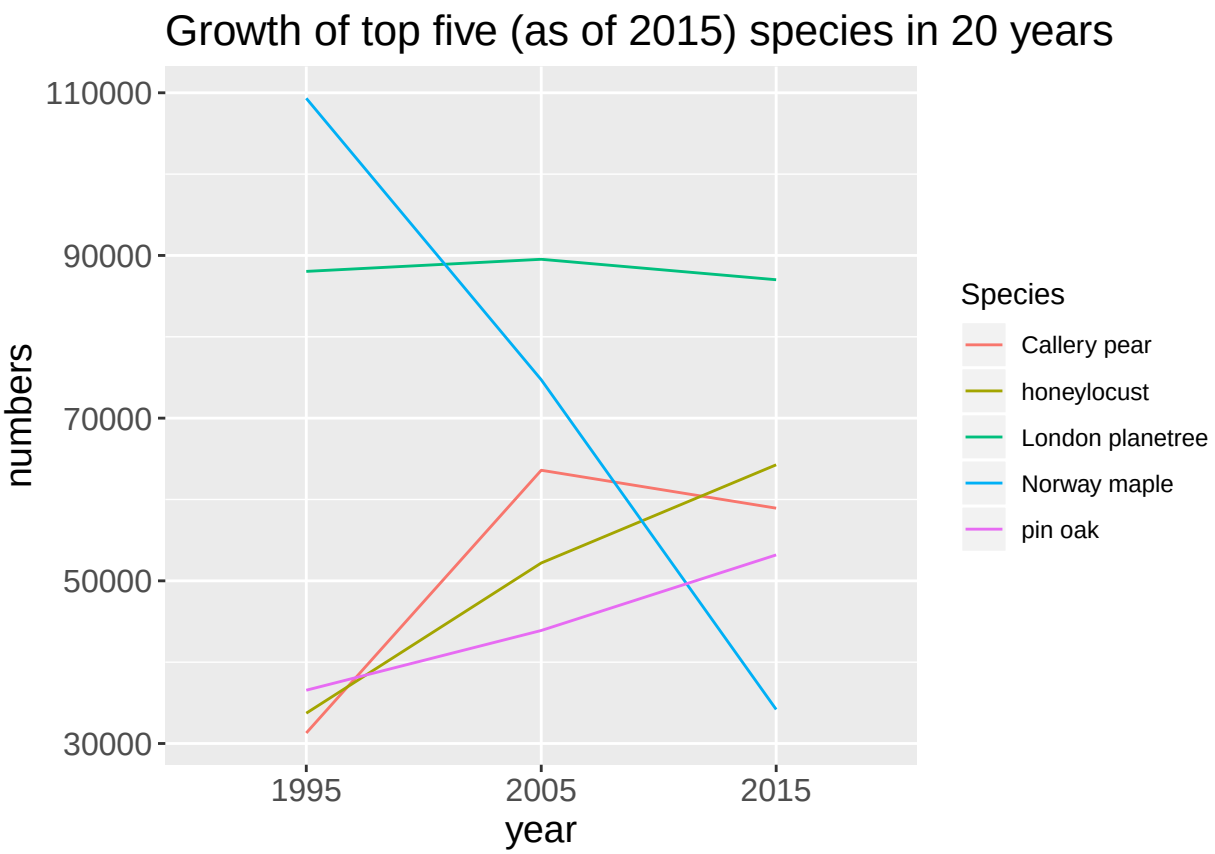
- Density of trees in NYC
- Growth of top five (as of 2015) species from 1995 to 2015
- Different health problems of trees

We will present the graphs and discuss our findings in the following pages.

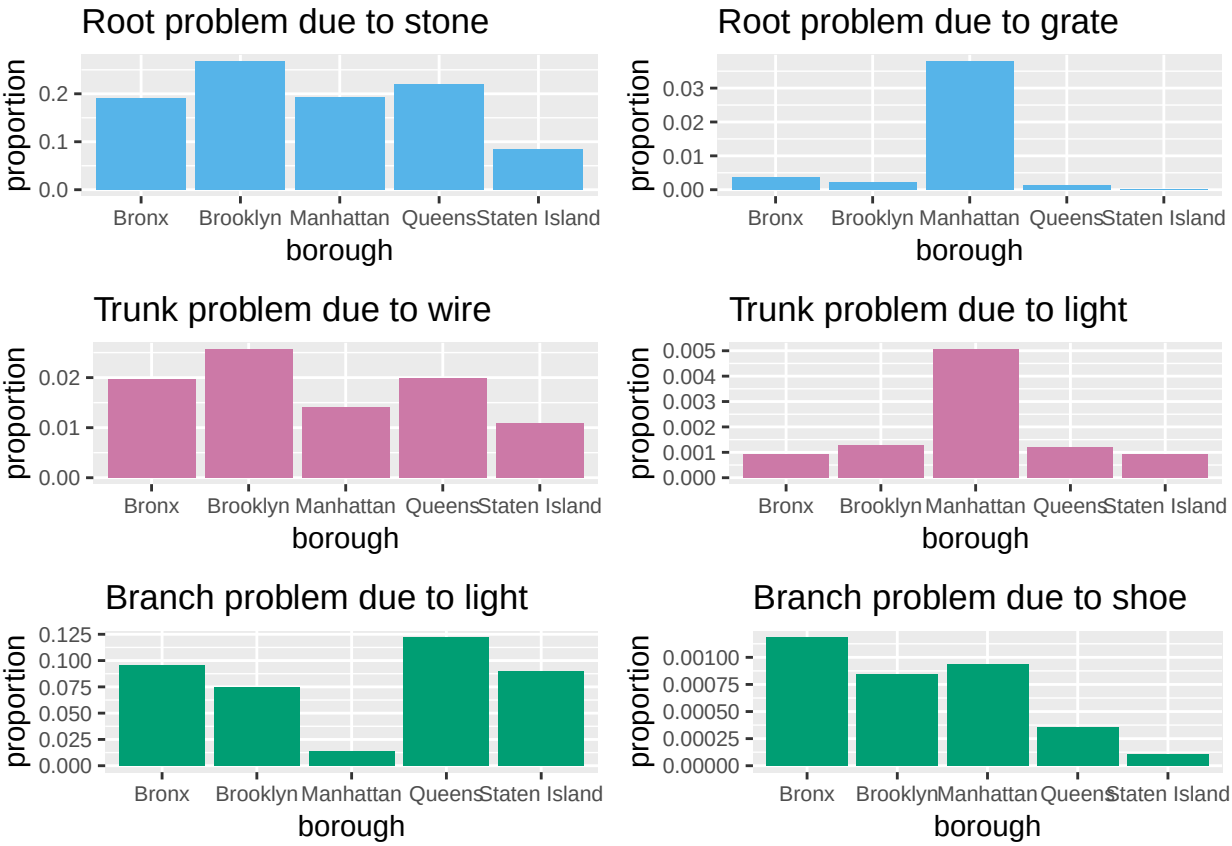
Density of trees (per square mile) in NYC



It can be seen from the spatial heat map that the higher percentage of trees mostly fall in the area around central park while less percentages of trees lie in Midtown Manhattan and even more less fall in areas in Staten Island. Overall, there are much more regions with deeper green than the lighter green showing that NYC has maintain a good ratio of green space. It is known to all that New York city is the most densely populated major city in the United States. Therefore, Midtown Manhattan, the largest central business district in the world with the majority of New York City's skyscrapers, has less ratio of green space compared with other regions.



In year 2015, the top five tree species in New York city are London planetree, honeylocust, Callery pear, pin oak and Norway maple. Each of five species presented a different growing trend in the past ten year. For example, the growing trend of London planetree is very stable. The most interesting is Morway maple. Knowing that the Norway maple is very invasive, the NYC government pulled seedlings of Norway maple from moist soil before they get too large to prevent them from destroying native ecosystems, causing trouble in yards and gardens, and creating visual blight. Therefore, the number of Norway maple decreased dramatically in the past ten years, varying from the first to the last.



From the plot above, we can observe many patterns. We can see that among the 6 problems, root problems caused by stone has a much higher percentage than other problem, around 20% of trees has this problems, while almost no trees has branch problem caused by shoe, and we can see that Manhattan has a much higher percentage of perblem in root problem caused by grate and truck problem caused by light, while a much lower problem proportion in brach problem caused by light. Also, we can see that Staten Island has a relatively lower proportion on almost every problem meaning that it has a good preservation and management of the trees.

Interactive component

Our interactive component is here. The plots will change based on the year and zipcode entered. Users maybe interested in the most common species that they live in. This is impossible to display with static graphs because there are too many zipcodes. One trend we want the users to discover is that the tree dbh distribution is shifting to the right which confirms the fact that trees are growing thicker.

Conclusion

After completing the project, we learn and practice our skills to clean, visualize and understand data. Especially, we learn from plotting density curve and species count for each zip code that if there is too much graphs to be plotted, using an interactive plot is a good idea. Otherwise, in our case, it is not practical to plot these 300 plots all in the R markdown file. Besides, we discovered the pros and cons for each type of graph, and learned to decide which graphs we should use. For example, we use a line graph to plot the changing of number of the species throughout these 20 years, and the line graph gives us a clear view of the changes better than for example, bar plot. Also, we found out that Github is a useful platform for groups work. When small changes happen, do a pull and push is a more convenient way than upload and download. But we did not utilize its power. Most of the time we just simply do downloads and uploads, and we will try to make more use of Github next time. This is a very meaningful project, we learn a lot from this experience.