Homework #3

For questions 1-4 in this problem set, we will work with a dataset on dogs of New York City, found here: https://project.wnyc.org/dogs-of-nyc/ (https://project.wnyc.org/dogs-of-nyc/)

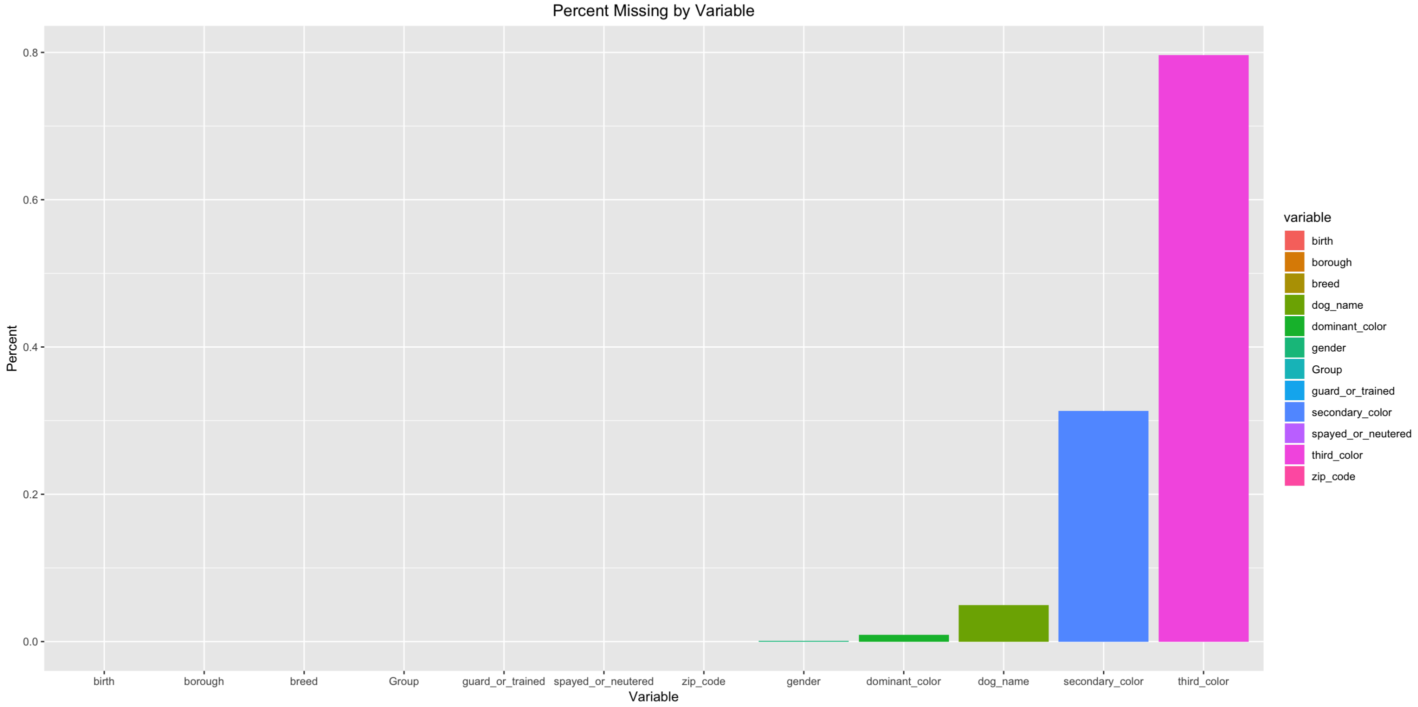
Background: The dataset is dated June 26, 2012. Although the data were originally produced by the NYC Department of Mental Health and Hygiene, it no longer seems to be available on any official NYC web site. (There is a 2016 dataset on dog licenses with different variables available here: https://data.cityofnewyork.us/Health/NYC-Dog-Licensing-Dataset/nu7n-tubp (https://data.cityofnewyork.us/Health/NYC-Dog-Licensing-Dataset/nu7n-tubp)). Also of note is the fact that this dataset has 81,542 observations. The same summer, the New York City Economic Development Corporation estimated that there were 600,000 dogs in New York City (source: https://blog.nycpooch.com/2012/08/28/how-many-dogs-live-in-new-york-city/ (https://blog.nycpooch.com/2012/08/28/how-many-dogs-live-in-new-york-city/)) Quite a difference! How many dogs were there really in 2012?!? Might be an interesting question to pursue for a final project, but for now we'll work with what we've got.

1. Missing Data

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(extracat)
library(stringr)
library(DataCombine)
library(scales)
library(data.table)
library(ggmosaic)

dog = fread("NYCdogs.csv",header = T, sep = ',')
dog[dog == "n/a"] = NA
```

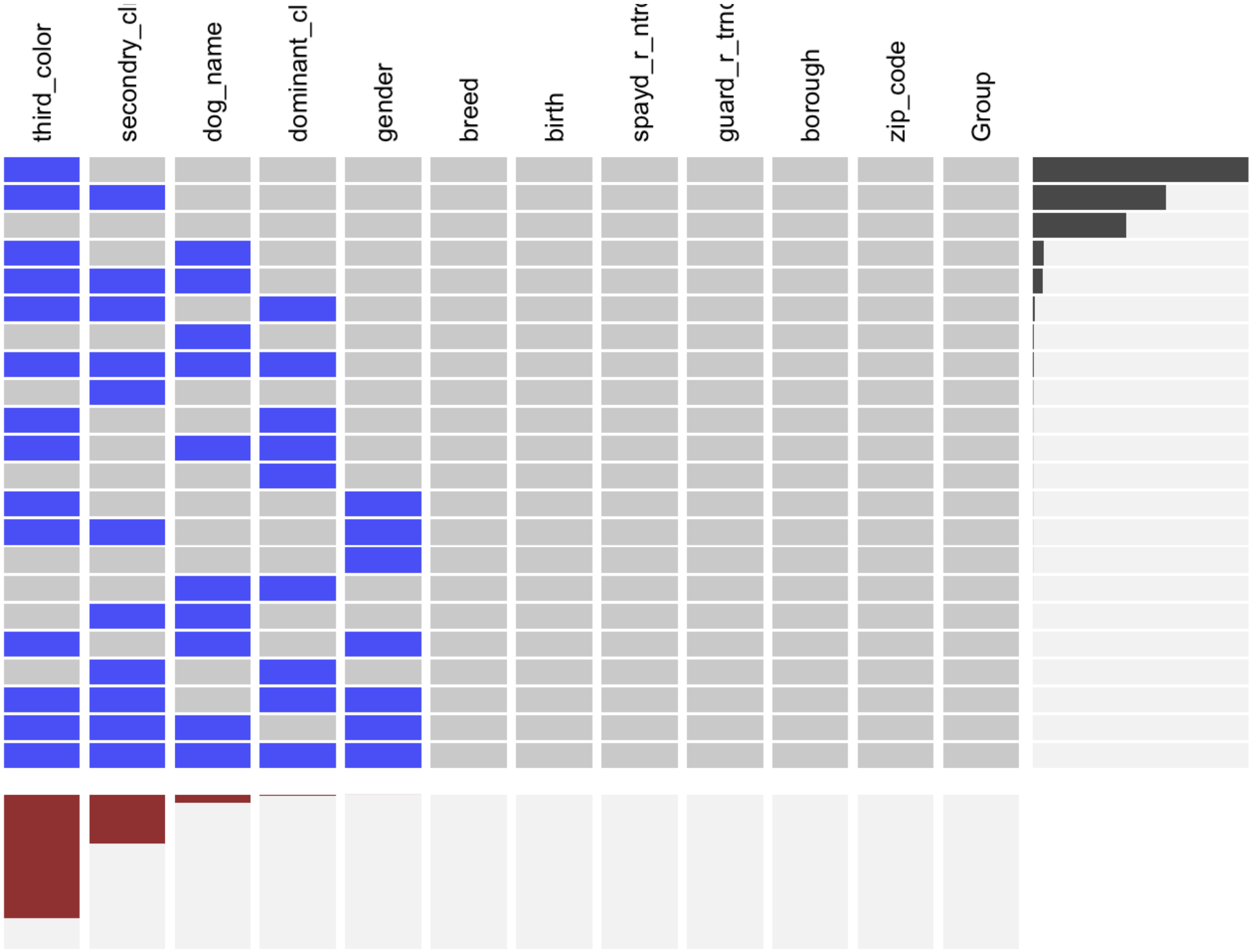
a. Create a bar chart showing percent missing by variable.

```
missing = colSums(is.na(dog)) / (colSums(!is.na(dog)) + colSums(is.na(dog)))
missing = cbind(names(dog),as.numeric(missing))
colnames(missing) = c("variable","count")
missing = data.frame(missing)
missing$count = as.numeric(as.character((missing$count)))

ggplot(missing,aes(x = reorder(variable,missing$count), y = missing$count, fill =
                     variable)) +
  geom_bar(stat = "identity") +
  xlab("Variable") +
  ylab("Percent") +
  ggtitle("Percent Missing by Variable") +
  theme(plot.title = element_text(hjust = 0.5))
```



b. Use the `extracat::visna()` to graph missing patterns. Interpret the graph.

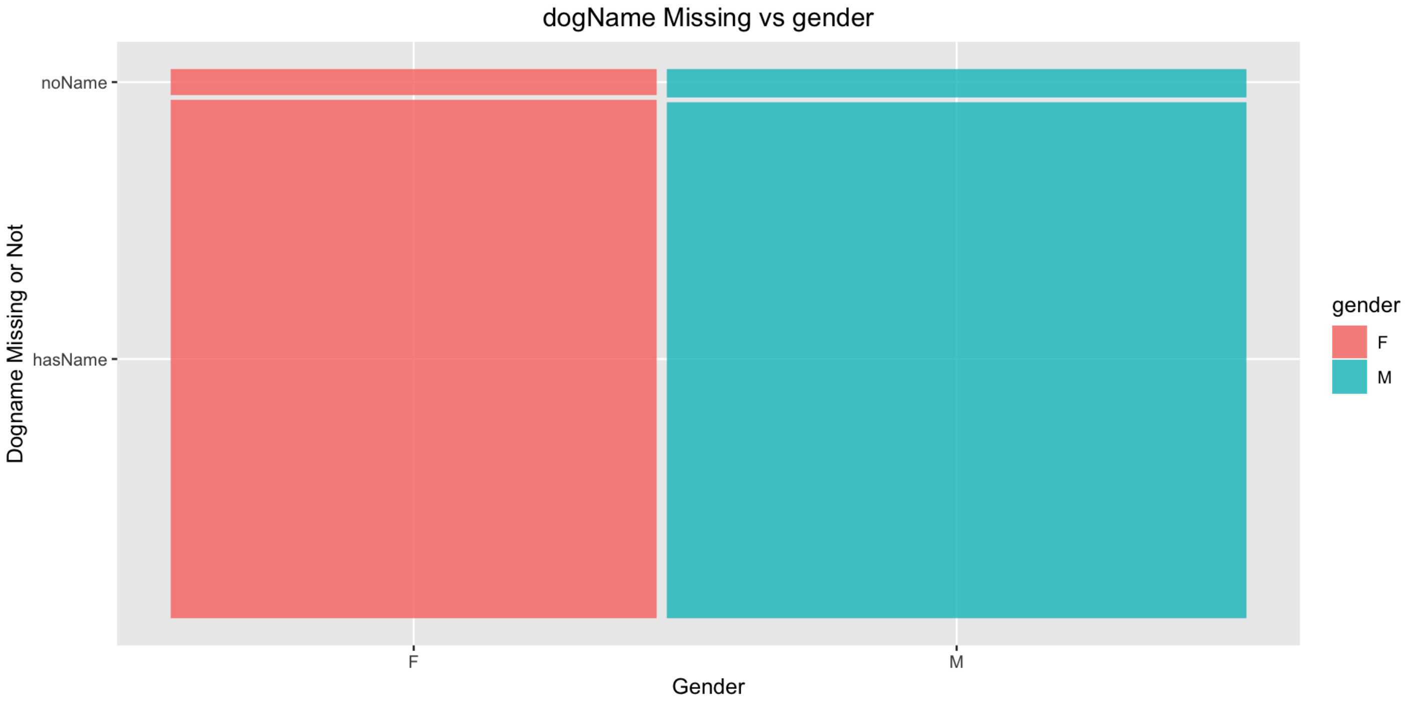
```
visna(dog, sort = "b")
```



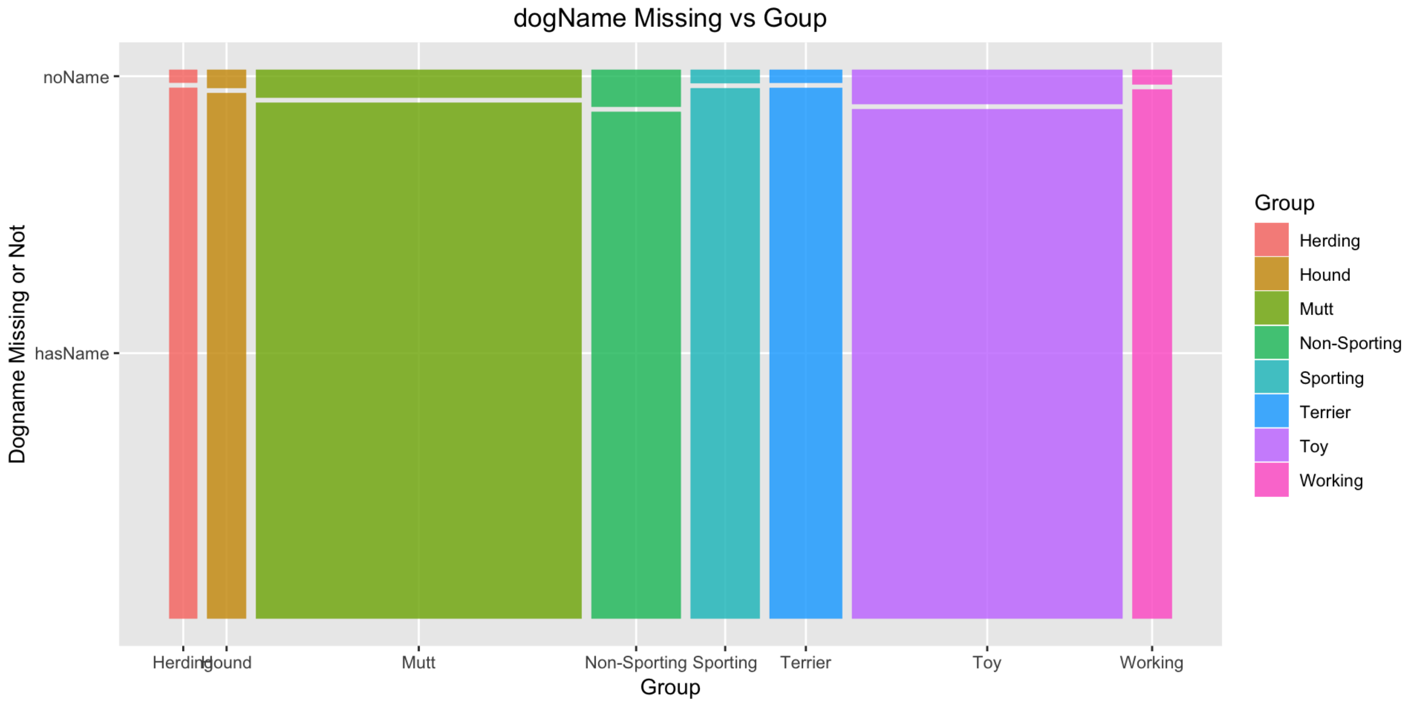
What I find is that the third color has the most missing data. The second, third most missing data are respectively secondary_color and dog name. Variables have missing values are thrid color,secondary color,dominant color, gender.

c. Do `dog_name` missing patterns appear to be associated with the *value* of `gender` , `Group` *or* `borough` ?

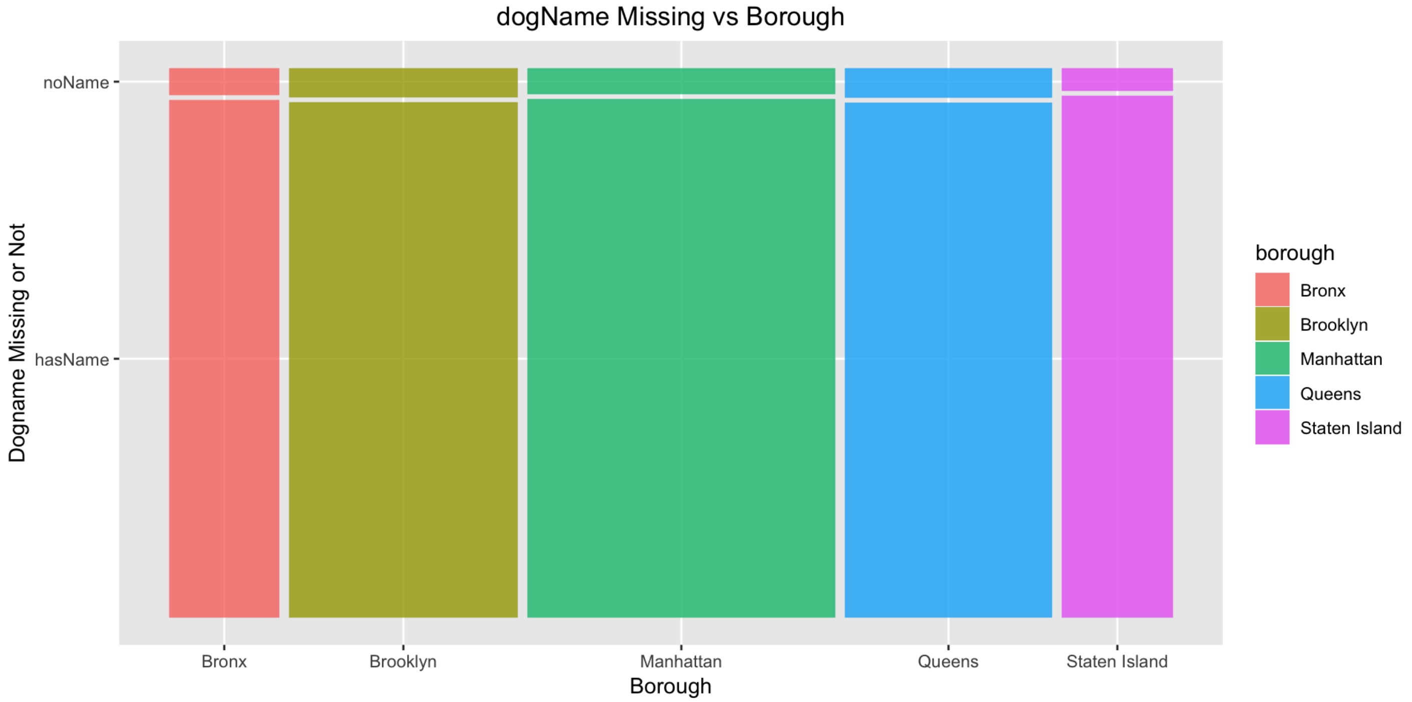
```
dog$dog_name[!is.na(dog$dog_name)] <- "hasName"
dog$dog_name[is.na(dog$dog_name)] <- "noName"

ggplot(dog) + geom_mosaic(aes(x = product(dog_name,gender),fill = gender),na.rm = T) +
  ggtitle("dogName Missing vs gender") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Dogname Missing or Not") +
  xlab("Gender")
```



```
ggplot(dog) + geom_mosaic(aes(x = product(dog_name,Group),fill = Group),na.rm = T) +
  ggtitle("dogName Missing vs Goup") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Dogname Missing or Not") +
  xlab("Group")
```



```
ggplot(dog) + geom_mosaic(aes(x = product(dog_name,borough),fill = borough),na.rm = T) +
  ggtitle("dogName Missing vs Borough") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Dogname Missing or Not") +
  xlab("Borough")
```



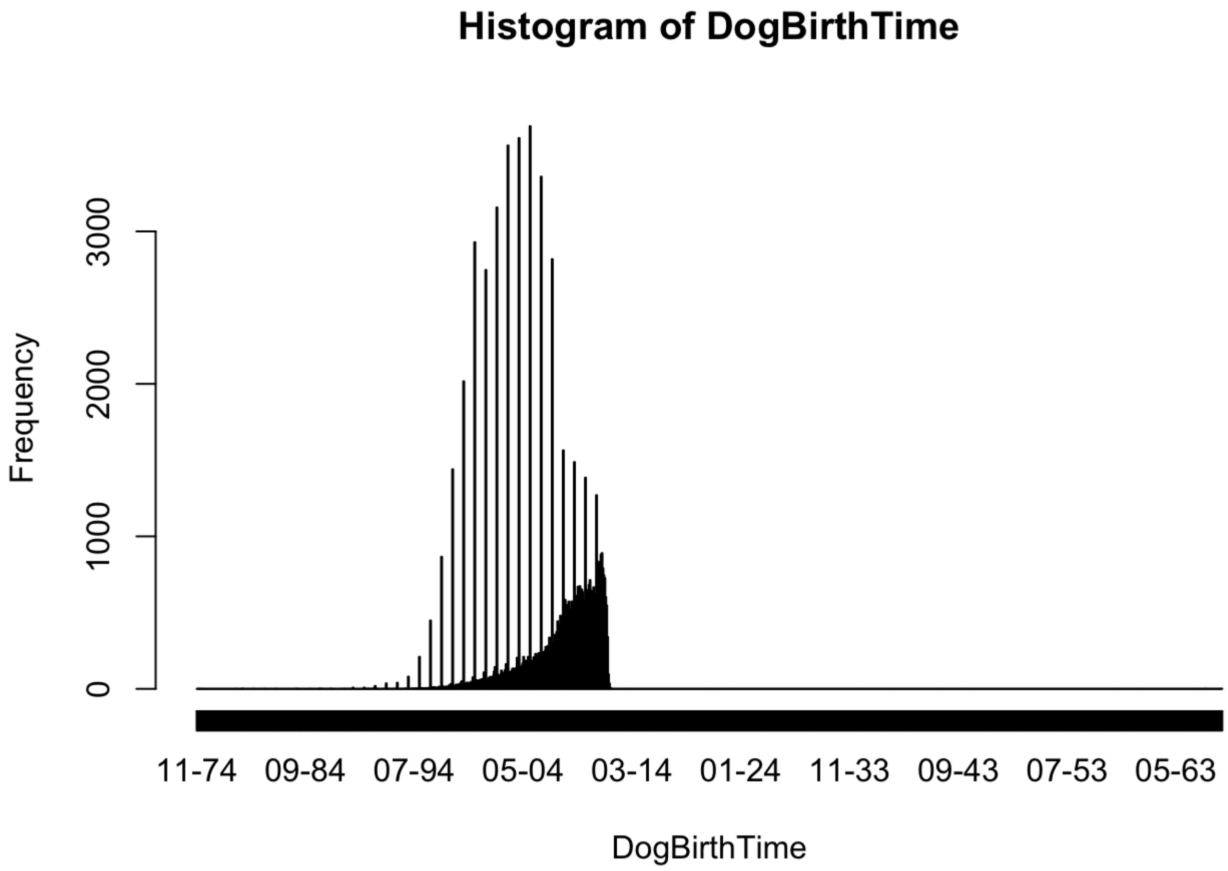
From the graph, we can see that dog name missing doesn't seem to be associated with gender. Neither does it associated with borough since for each of those variables, the proportion between has name and not has name seems the same. Dog name could have some association with Group as what we see from the mosiac plot for non-sporing group, there are more dog name missing

2. Dates

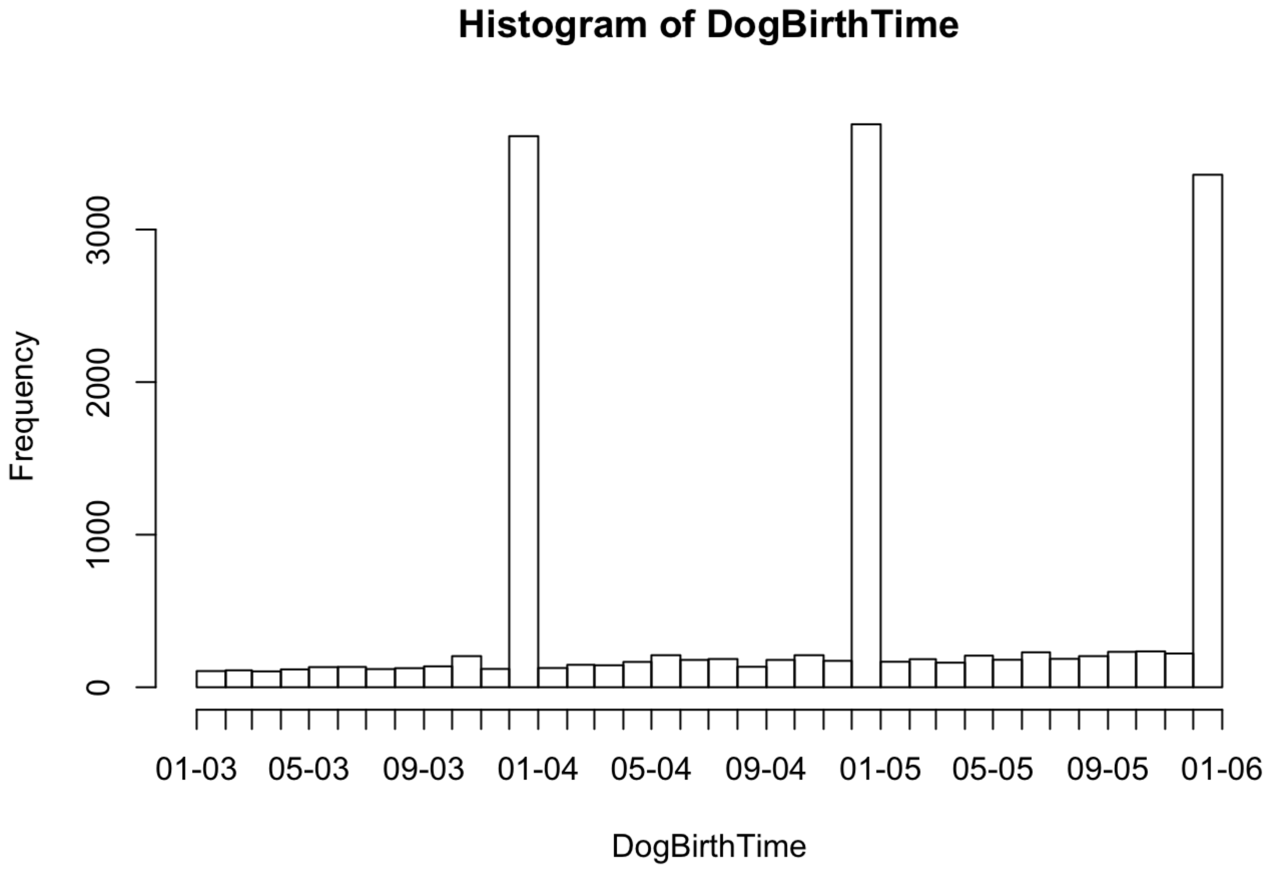
a. Convert the `birth` column of the NYC dogs dataset to `Date` class (use “01” for the day since it’s not provided). Create a frequency histogram of birthdates with a one-month binwidth. (Hint: don’t forget about base R.) What do you observe? Provide a reasonable hypothesis for the prominent pattern in the graph.

```
dog$Month = str_extract(dog$birth,"[A-Z]+[a-z]{1,4}")
dog$Year = str_extract(dog$birth,"[0-9]{1,2}")
dog$MY = paste("01",dog$Month,dog$Year,sep = "-")
dog$birth = as.Date(dog$MY,format = "%d-%b-%y")
DogBirthTime = as.POSIXct(dog$birth)
```

```
hist(DogBirthTime,"months", format = "%m-%y", freq = T)
```



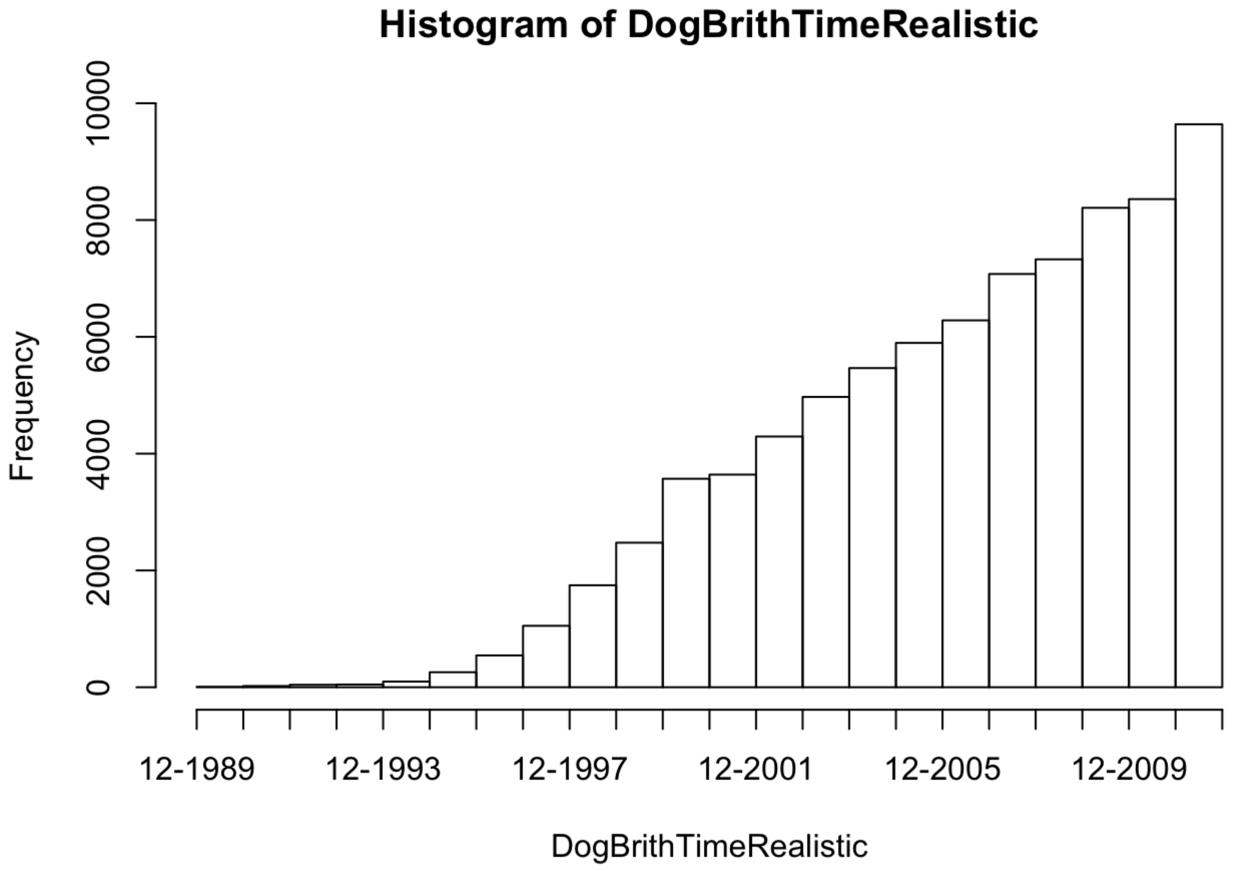
```
DogBirthTime <- subset(DogBirthTime, DogBirthTime > "2003-01-01")
DogBirthTime <- subset(DogBirthTime, DogBirthTime < "2006-01-01")
hist(DogBirthTime,"months", format = "%m-%y", freq = T, right = F)
```



The first thing I see is that the year clearly exceed the current year, these are impossible dates in the data. This happen probabily due to the date conversion method I use. I could also possible casued by the quality of data. There are exceptional high number of birth occur at January of each year. The reason for this could be that when people are asked about birth date of their dogs, they cannot remember and they just give January as result.

b. Redraw the frequency histogram with impossible values removed and a more reasonable binwidth.

```
dog1 <- subset(dog, dog$birth < "2012-01-01")
dog1 <- subset(dog1, dog1$birth > "1990-01-01")
DogBrithTimeRealistic = as.POSIXct(dog1$birth)
hist(DogBrithTimeRealistic,"years", format = "%m-%Y", freq = T)
```

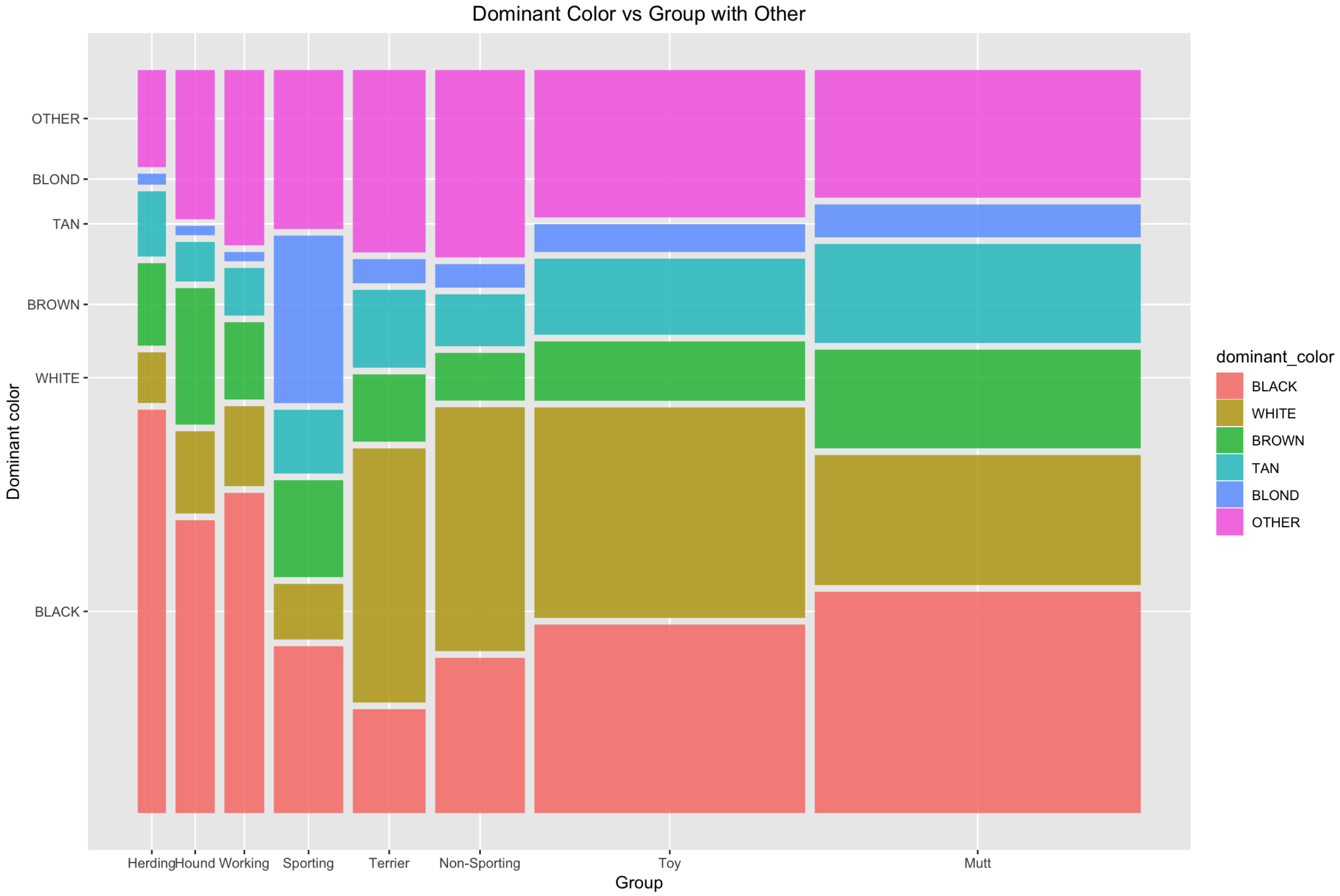


I subset the data from 1990-01-01 to 2012-01-01. Dogs normally don’t live longer than 20 years, it doesn’t make much sense to include dogs birth 20 years earlier than 2012.It also doesn’t make sense to include any birth date of dog after 2012.

3. Mosaic plots

a. Create a mosaic plot to see if `dominant_color` depends on `Group`. Use only the top 5 dominant colors; group the rest into an “OTHER” category. The last split should be the dependent variable and it should be horizontal. Sort each variable by frequency, with the exception of “OTHER”, which should be the last category for dominant color. The labeling should be clear enough to identify what’s what; it doesn’t have to be perfect. Do the variables appear to be associated? Briefly describe.

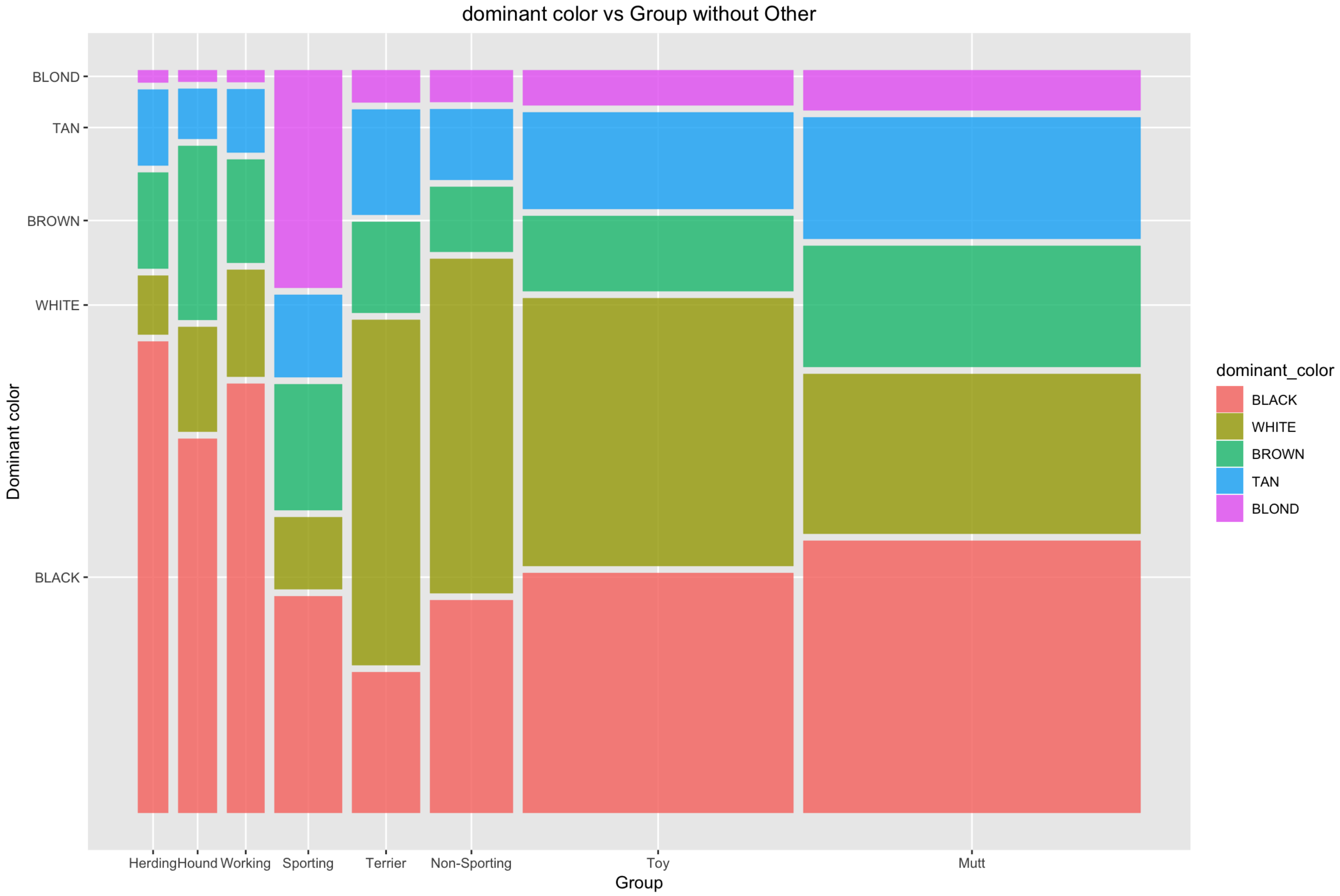
```
dogs = fread("NYCdogs.csv",header = T, sep = ',')
dogs[dogs == "n/a"] = NA
c = names(sort(summary(factor(dogs$dominant_color)),decreasing = TRUE)[1:5])
dogs$dominant_color[!dogs$dominant_color %in% c]<-"OTHER"
dogs <- within(dogs, dominant_color <- factor(dominant_color, levels = c(c,"OTHER")))
dogs <- within(dogs, Group <- factor(Group, levels = names(sort(summary(factor(dogs$Grou
p))))))
ggplot(dogs) + geom_mosaic(aes(x = product(dogs$dominant_color,Group),fill = dominant_co
lor),na.rm = T) +
  ggtitle("Dominant Color vs Group with Other") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Group") +
  ylab("Dominant color")
```



The dominant_color five color seems associated with the group. For herding group, there more black. For sporting, there are more blond dogs. For terrier group, there appear to be more white dogs. There are many similar patterns in each category of Group which suggest that dogs colors are associated with group.

b. Redraw with the “OTHER” category filtered out. Do the results change? How should one decide whether it’s necessary or not to include an “OTHER” category?

```
dogs = dogs[dogs$dominant_color != "OTHER"]
dogs <- within(dogs,dominant_color <- factor(dominant_color,levels = names(sort(table(fa
ctor(dominant_color)),decreasing = T))))
ggplot(dogs) + geom_mosaic(aes(x = product(dominant_color,Group),fill = dominant_color)
                           ,na.rm = T) +
  ggtitle("dominant color vs Group without Other") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Group") +
  ylab("Dominant color")
```

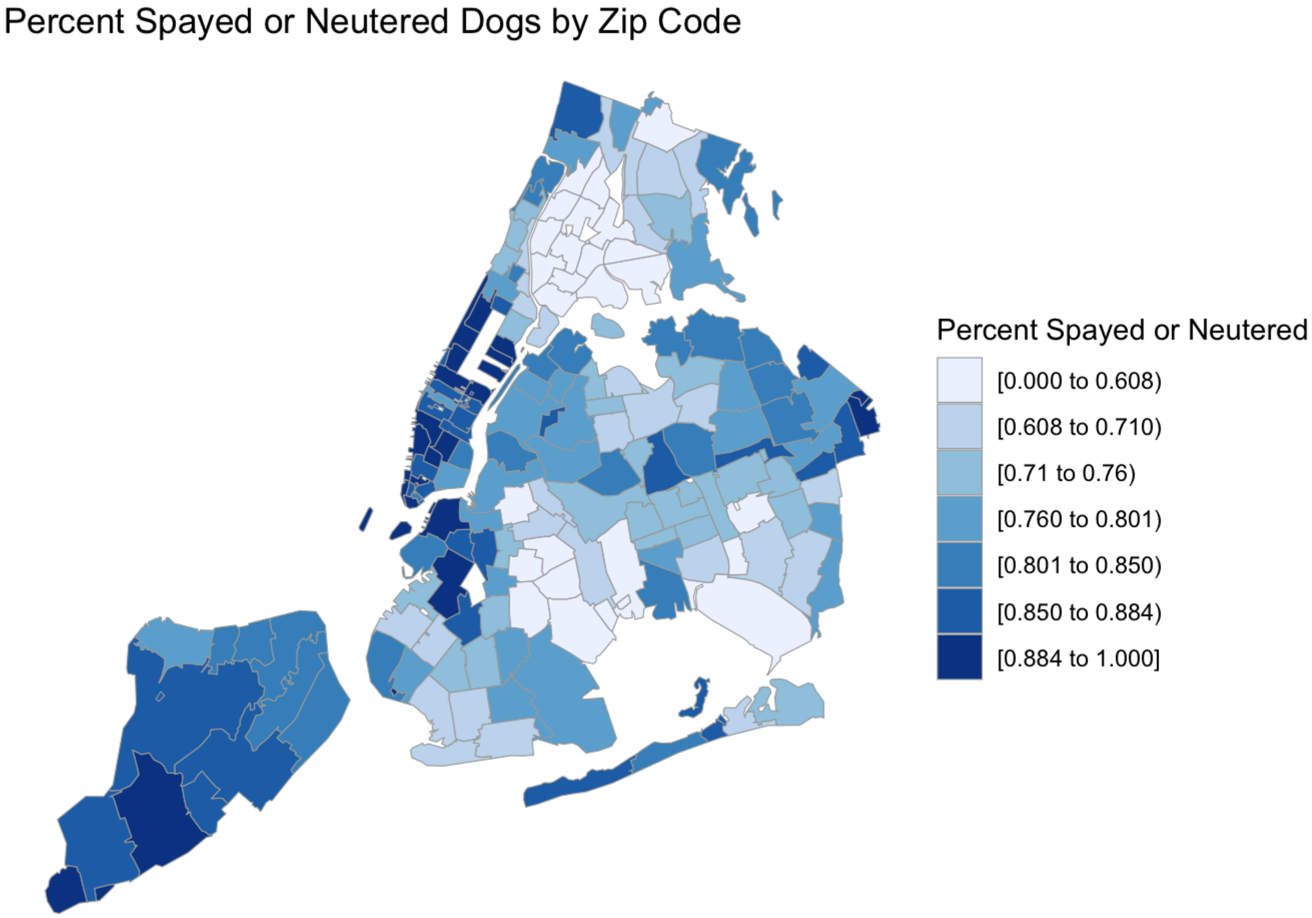


The result doesn't change when comparing within the top five dominant color. However, after removing other from the graph, we can capture the difference more accurately. When trying to analyze the entire dominant_color dataset, we should add OTHER in so that I can get a sense of how does the sum of the rest of dominat color besides the first five compare. However, when try to comprehand a relationship within the five dominant color, one should remove OTHER from the graph

4. Maps

Draw a spatial heat map of the percent spayed or neutered dogs by zip code. What patterns do you notice?

```
library(choroplethr)
library(choroplethrMaps)
require(choroplethrZip)

zip = levels(factor(dogs$zip_code))
percent = c()
for(i in 1:length(zip)){
  percentadd = summary(factor(dogs$spayed_or_neutered[dogs$zip_code == zip[i]]))["Yes"]/
sum(summary(factor(dogs$spayed_or_neutered[dogs$zip_code == zip[i]])))
  percent = cbind(percent,percentadd)
}
percent[is.na(percent)] <- 0
data = cbind(zip,t(percent))
df <- data.frame(data)
colnames(df) <- c("region","value")
rownames(df) <- c()

data("zip.regions")
df$region <- as.character(df$region)
df$value <- as.numeric(as.character((df$value)))
df = df[df$region %in% zip.regions$region,]
zip_choropleth(df, zip_zoom = df$region,title = "Percent Spayed or Neutered Dogs by Zip
 Code",legend=c("Percent Spayed or Neutered"))
```

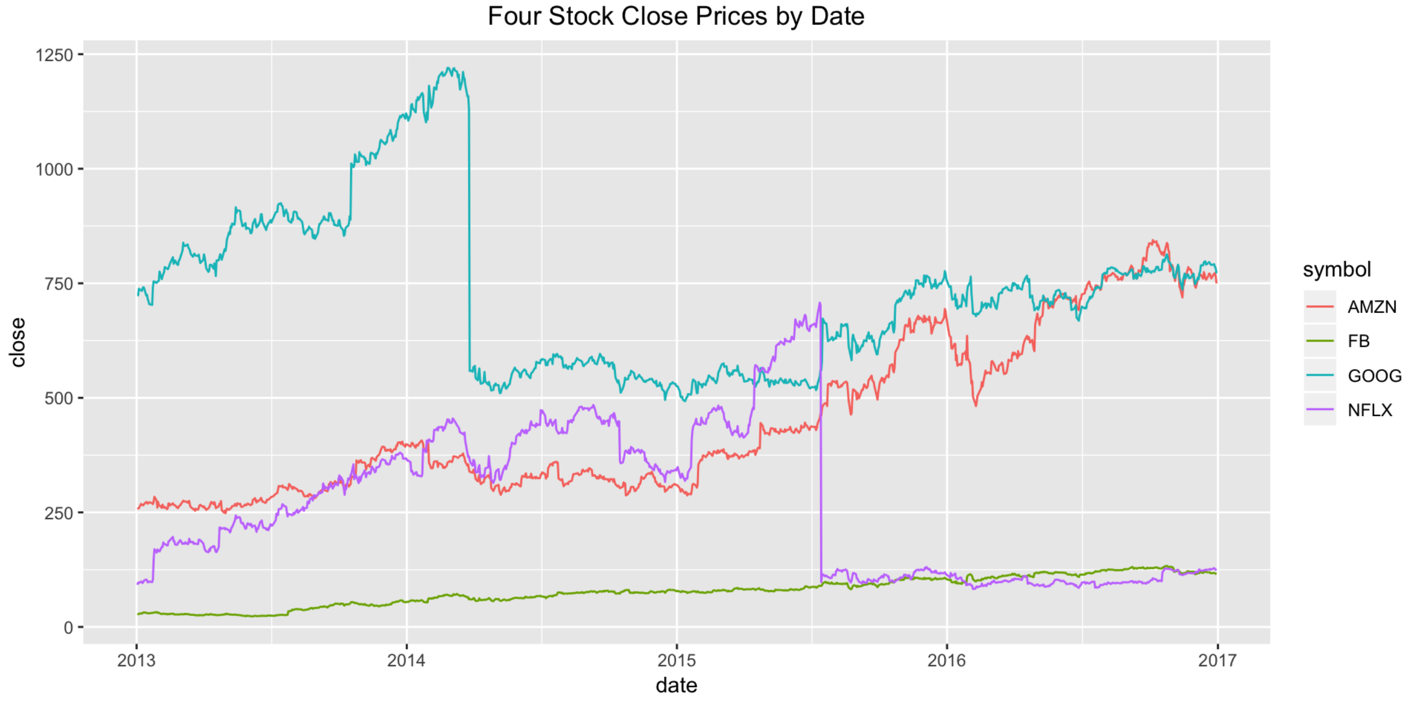


What I can observe from this graph is that in the downtown of manhattan, there are clearly more dog being spayed. On the contrary, less dogs are spayed in uptown of manhattan. In addtiont, Staten Island also has a high percentage of dog being spayed.

5. Time Series

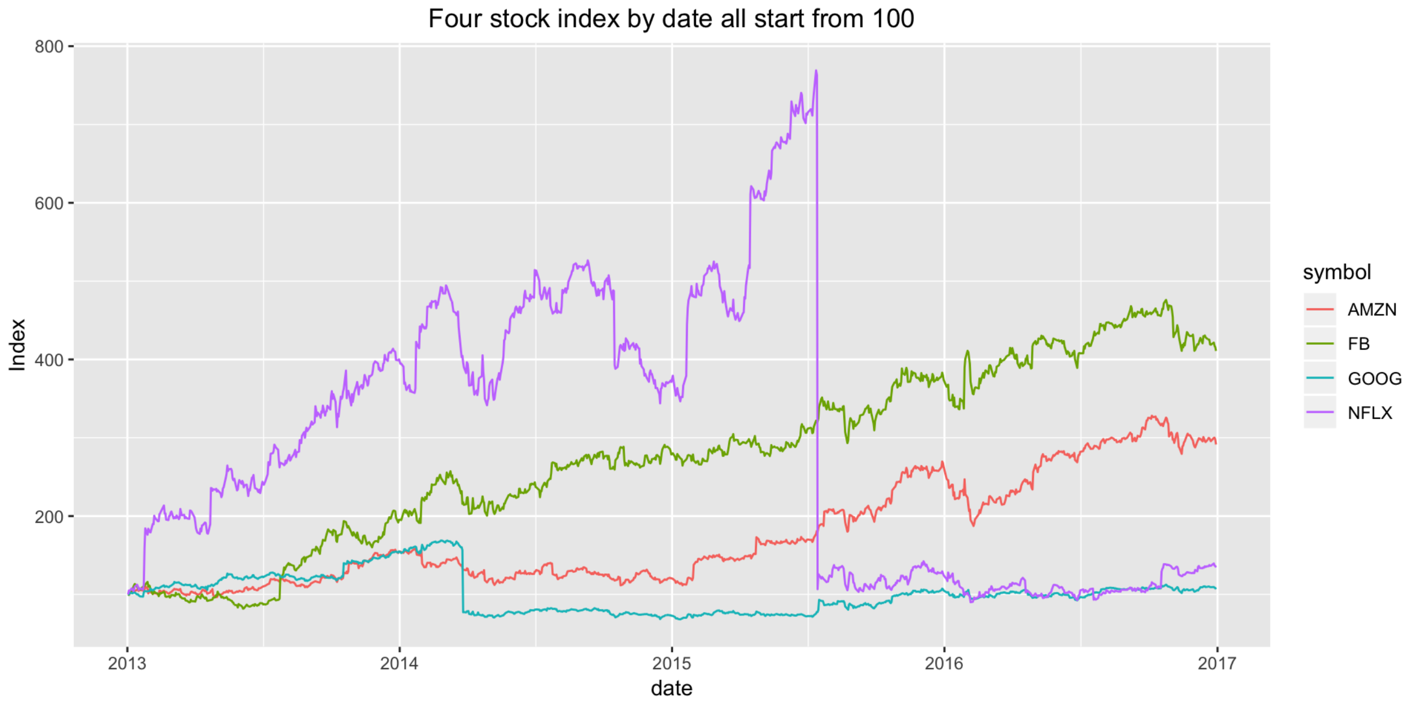
a. Use the `tidyquant` package to collect information on four tech stocks of your choosing. Create a multiple line chart of the closing prices of the four stocks on the same graph, showing each stock in a different color.

```
library(tidyquant)
data('FANG')

ggplot(FANG,aes(x = date, y = close, color = symbol)) + geom_line() +
  ggtitle("Four Stock Close Prices by Date") +
  theme(plot.title = element_text(hjust = 0.5))
```



b. Transform the data so each stock begins at 100 and replot. Choose a starting date for which you have data on all of the stocks. Do you learn anything new that wasn't visible in (a)?

```
FANG$close[FANG$symbol == "AMZN"] <- FANG$close[FANG$symbol == "AMZN"] / FANG$close[FANG$symbol == "AMZN"][1]*100
FANG$close[FANG$symbol == "FB"] <- FANG$close[FANG$symbol == "FB"] / FANG$close[FANG$symbol == "FB"][1]*100
FANG$close[FANG$symbol == "GOOG"] <- FANG$close[FANG$symbol == "GOOG"] / FANG$close[FANG$symbol == "GOOG"][1]*100
FANG$close[FANG$symbol == "NFLX"] <- FANG$close[FANG$symbol == "NFLX"] / FANG$close[FANG$symbol == "NFLX"][1]*100

ggplot(FANG,aes(x = date, y = close, color = symbol)) + geom_line() +
  ggtitle("Four stock index by date all start from 100") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Index")
```



From this graph, I can acquire information of percent increase for each stock compare to their initial price. Nflx increased almost 700 percent from 2013 to mid 2015 which has the highest percent of increase compare to the rest of three stocks. Google despite a rapid drop in stock price at mid 2014, but the percent dercrease is not that much since the inital starting price for google is much higher. Amazon and Facebook exhibit steady increase of stock price over time. Facebook depsite has lower price than Amazon has higher rate of increase.

6. Presentation

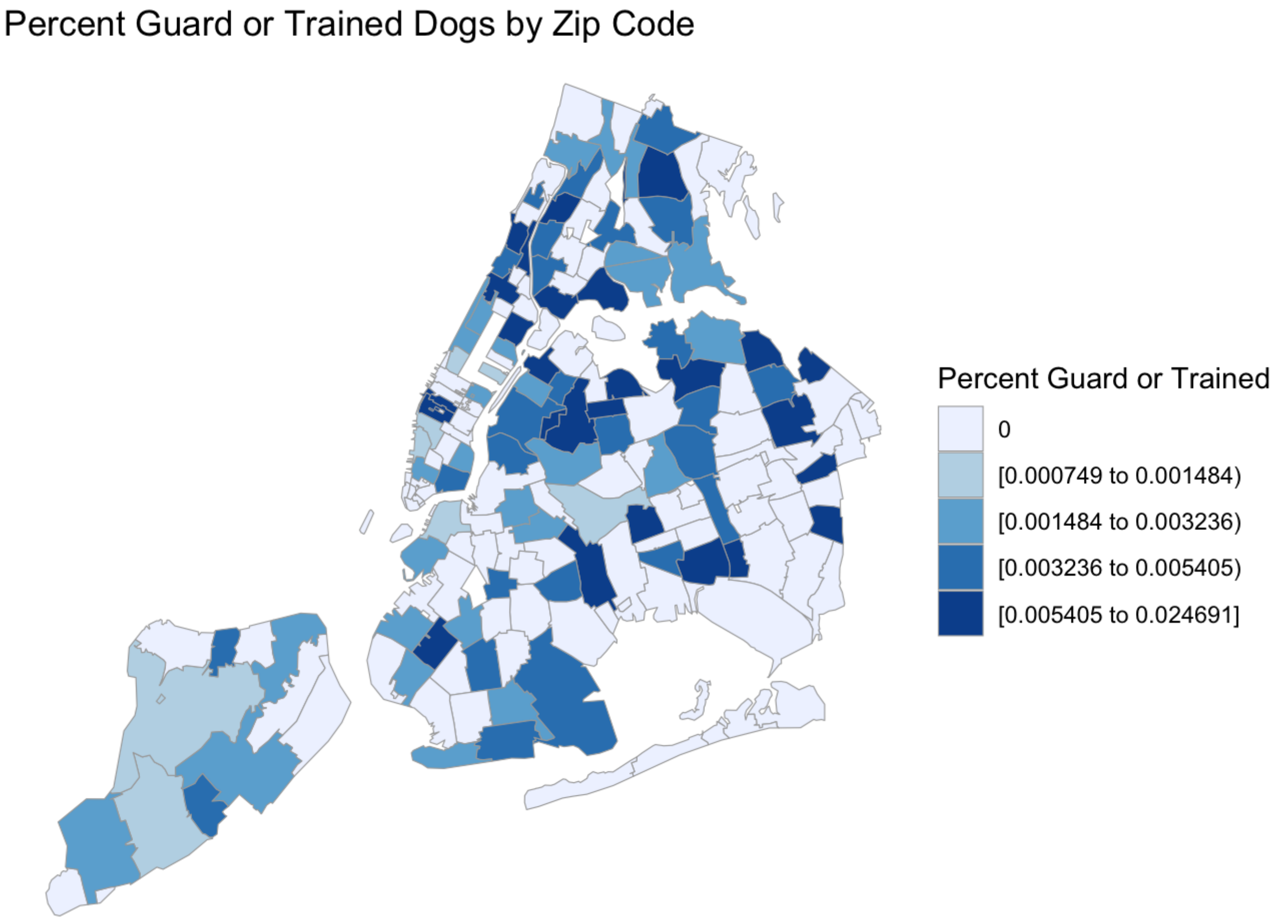
Imagine that you have been asked to create a graph from the Dogs of NYC dataset that will be presented to a very important person (or people). The stakes are high.

a. Who is the audience? (Mayor DeBlasio, a real estate developer, the voters, the City Council, the CEO of Purina…) Let the audience be Mayor Deblasio

b. What is the main point you hope someone will take away from the graph? First let the Mayor know that majority of the dogs in New York City is not trainned.To reduce potential home invasion,funds should be issued to initialized dog training programs.In addition to that,the dog tranining facilities should be build in different places in New York City, so that the difference of the percent of trained dog in each zip region is reduced. I draw a sptial heatmap on the percent of dog trained to demonstrate the above two points.

c. Present the graph, cleaned up to the standards of “presentation style.” Pay attention to choice of graph type, if and how the data will be summarized, if and how the data will be subsetted, title, axis labels, axis breaks, axis tick mark labels, color, gridlines, and any other relevant features.

```
percentTrained = c()
for(i in 1:length(zip)){
  percentTrainedAdd = summary(factor(dogs$guard_or_trained[dogs$zip_code == zip[i]]))["Y
es"]/sum(summary(factor(dogs$spayed_or_neutered[dogs$zip_code == zip[i]])))
  percentTrained = cbind(percentTrained,percentTrainedAdd)
}
percentTrained[is.na(percentTrained)] <- 0
data1 = cbind(zip,t(percentTrained))
df1 <- data.frame(data1)
colnames(df1) <- c("region","value")
rownames(df1) <- c()
df1 <- df1[df1$region %in% zip.regions$region,]
df1$region <- as.character(df1$region)
df1$value <- as.numeric(as.character((df1$value)))
zip_choropleth(df1, zip_zoom = df1$region,title = "Percent Guard or Trained Dogs by Zip
 Code",legend=c("Percent Guard or Trained"))
```


Percent Guard or Trained Dogs by Zip Code
Percent Guard or Trained
0
[0.000749 to 0.001484)
[0.001484 to 0.003236)
[0.003236 to 0.005405)
[0.005405 to 0.024691]