Homework #3

Hrishikesh Telang(hnt2107)

10/30/2018

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/)

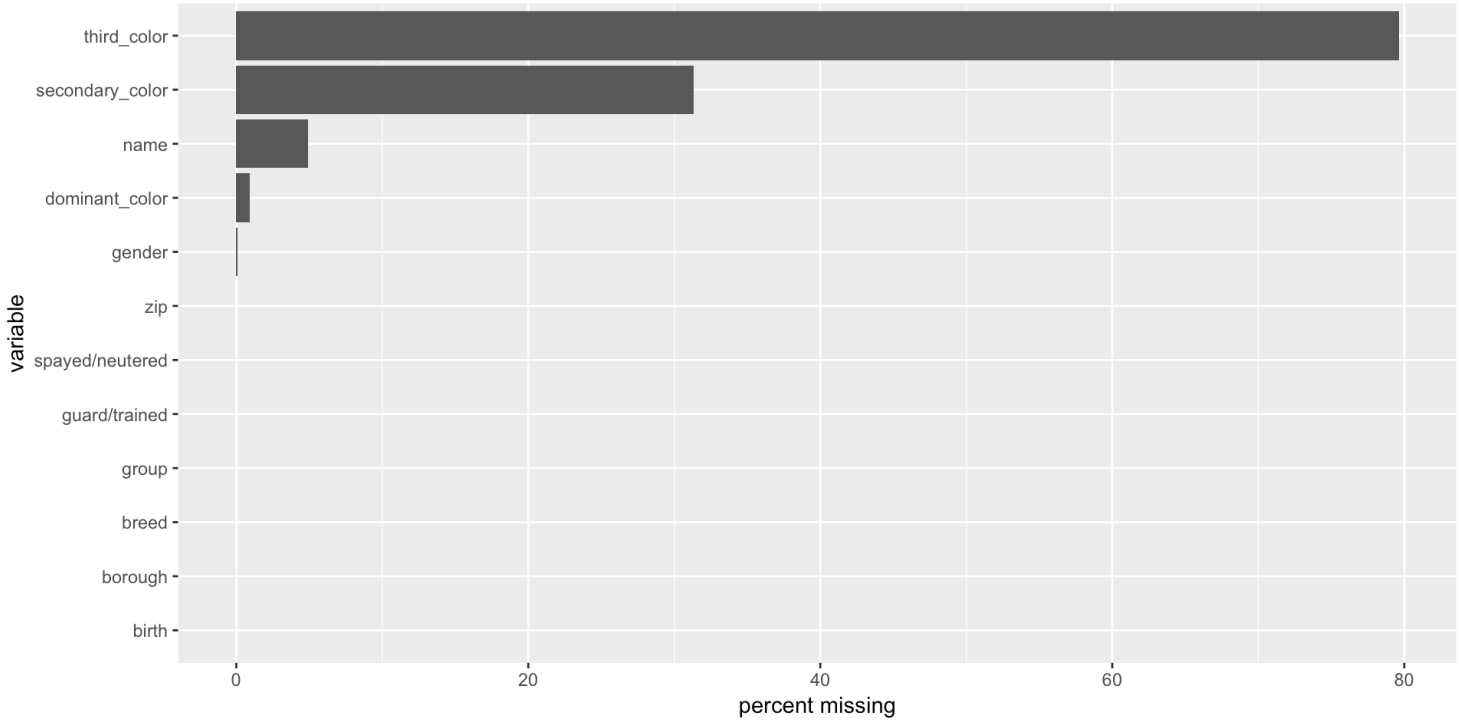
Please use the "NYCdogs.csv" version found in Files/Data folder on CourseWorks, which includes a Group column. If you already did some of the questions that didn't require the Group column, you do not have to redo them.

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

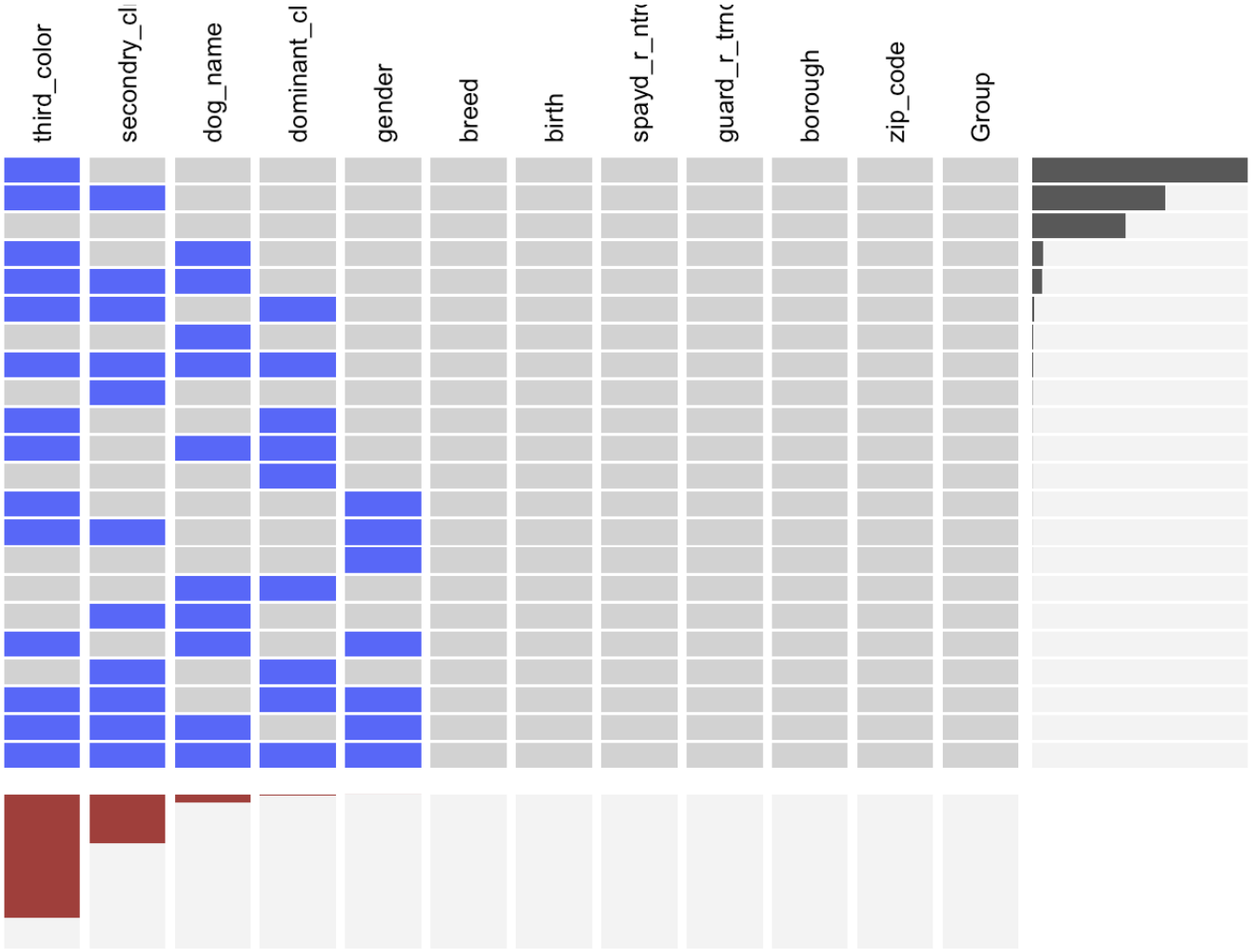
a. Create a bar chart showing percent missing by variable.

```
library(tidyverse)
library(dplyr)
library(plotly)
library(ggplot2)
df<-read_csv('/Users/hrishikeshtelang/Downloads/NYCdogs.csv')
per<-data.frame((colSums(is.na(df))%>%sort(decreasing=FALSE)))
colnames(per)<-("count")
p<-per%>%mutate(count=(count/81542)*100)

variables=c("breed","birth","spayed/neutered","guard/trained","borough","zip","group"
,"gender","dominant_color","name","secondary_color","third_color")
p<-cbind(p,variables)
l<-ggplot(p,aes(x=reorder(variables,count),y=count))+geom_bar(stat="identity")+xlab("
variable")+ylab("percent missing")
l+coord_flip()
```



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

```
library(extracat)
visna(df,sort='b')
```



1.The variable third_color has by far the most missing values compared to any other variable, followed by secondary_color.

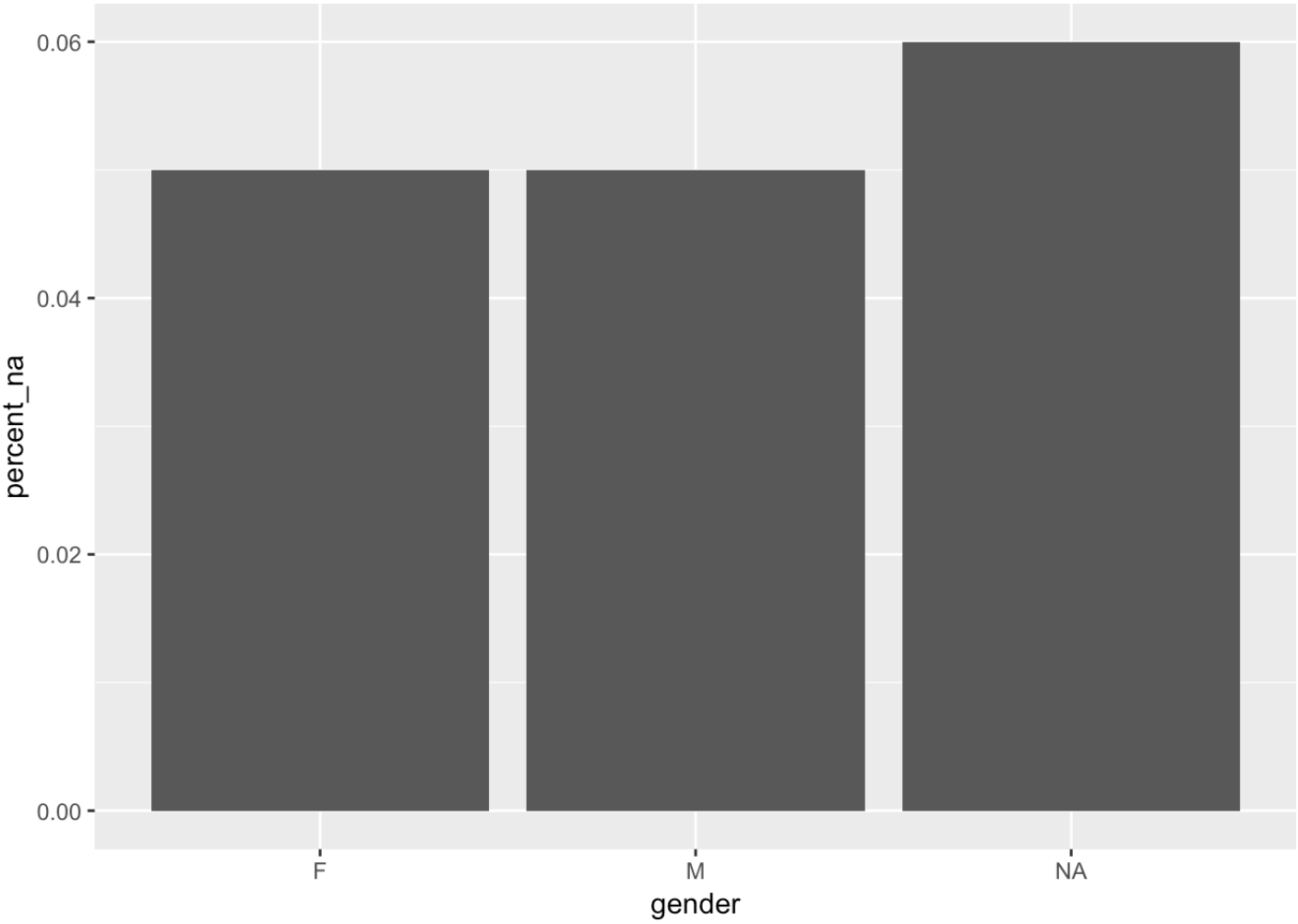
2.Only 5 variables have missing values

3.Rows having no missing values are the third most frequently occuring pattern

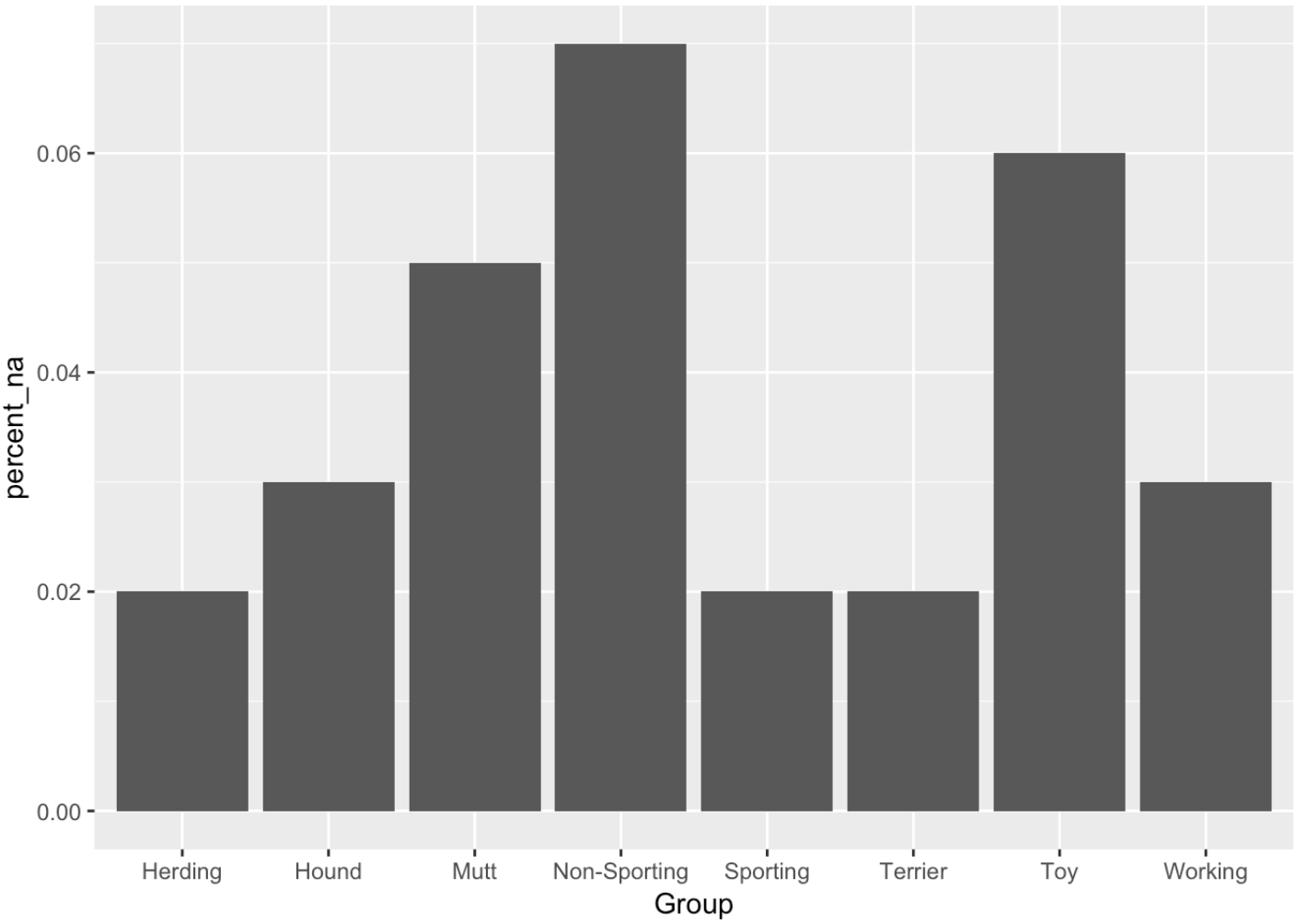
4.Very few rows have all 5 of these variables missing

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

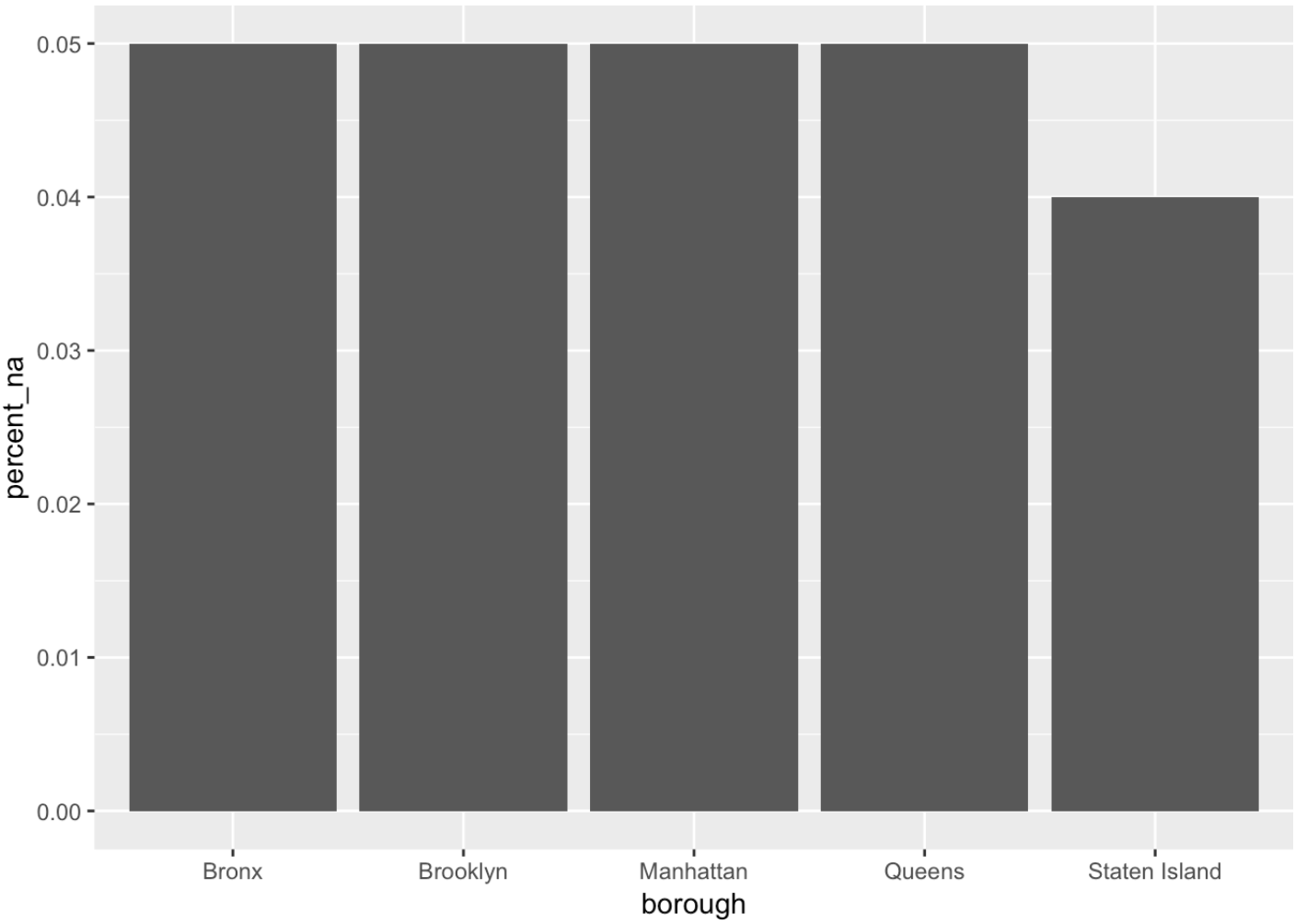
```
gen<-df %>%group_by(gender)%>%summarize(num_tot = n(), num_na = sum(is.na(`dog_name`)
)) %>% mutate(percent_na = round(num_na/num_tot, 2)) %>% arrange(-percent_na)

ggplot(gen,aes(x=gender,y=percent_na))+geom_bar(stat="identity")
```




```
gr<-df %>%group_by(Group)%>%summarize(num_tot = n(), num_na = sum(is.na(`dog_name`)))
%>% mutate(percent_na = round(num_na/num_tot, 2)) %>% arrange(-percent_na)

ggplot(gr,aes(x=Group,y=percent_na))+geom_bar(stat="identity")
```




```
bor<-df %>%group_by(borough)%>%summarize(num_tot = n(), num_na = sum(is.na(`dog_name`
))) %>% mutate(percent_na = round(num_na/num_tot, 2)) %>% arrange(-percent_na)

ggplot(bor,aes(x=borough,y=percent_na))+geom_bar(stat="identity")
```



1.While the graph does show a pattern for missing values of dog names and gender, it is very uncommon.

2.There appears to be almost an equal split amongst number of missing name values by gender.

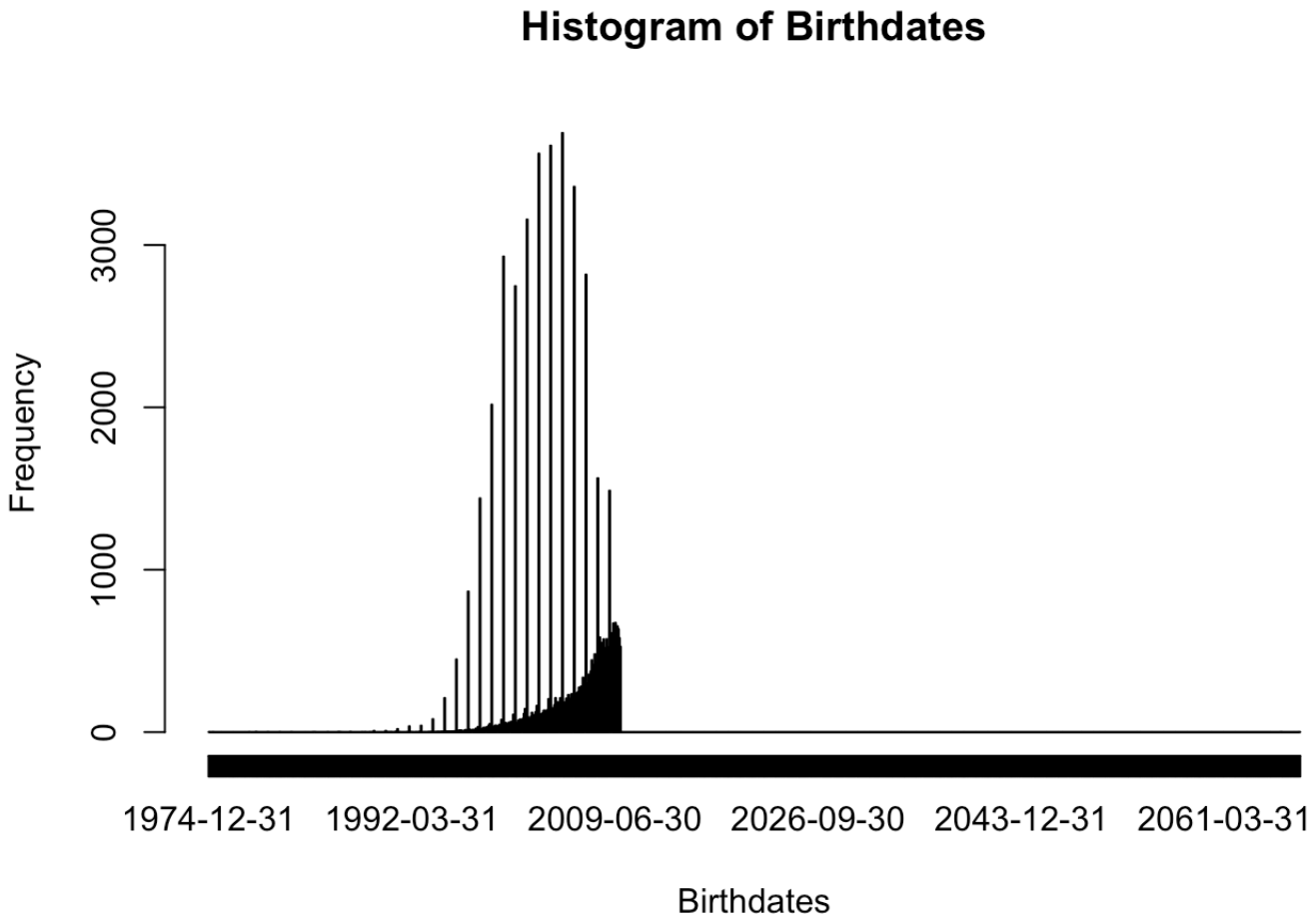
3.When analysing missing name values by Group, we notice that the Non-sporting category has the highest percentage of missing values followed by Toy and Mutt. There is a clear distribution of missing name values by group.

4.There appears to be no special pattern in missing data by borough, with all boroughs having almost equal percentages.

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.

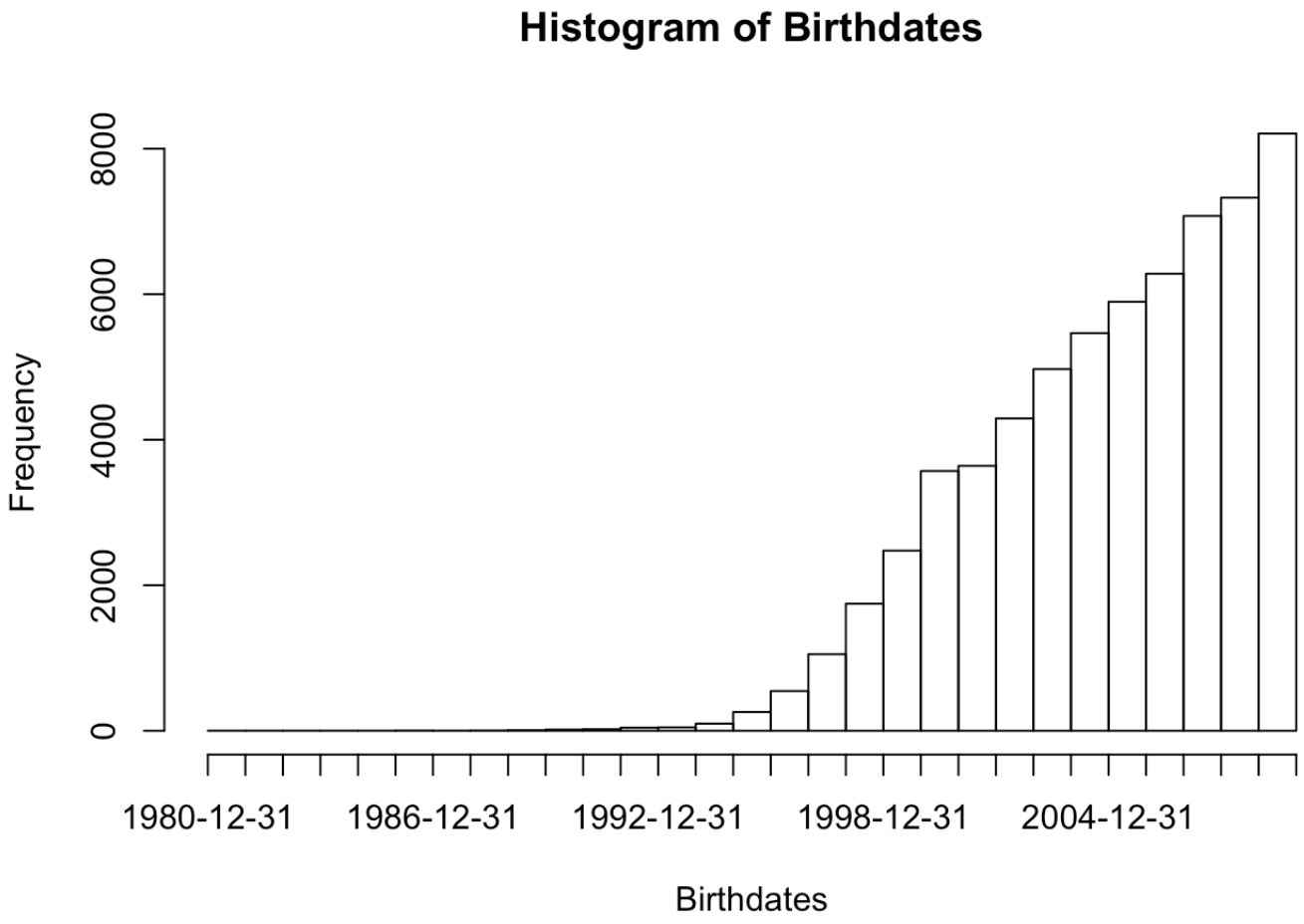
```
library(data.table)
library(lubridate)
new<-df%>%mutate(date=paste(birth,"01",sep="-"))

new<-new%>%mutate(date=ifelse(date%like%"^[1-9]",paste("0",date,sep=""),date))

new$date = parse_date_time(new$date, c("myd","ymd"))

hist(new$date,"months",xlab="Birthdates",freq=TRUE)
```



1. When the binwidth is set to months, there is an unusual pattern obtained which consists of odd spikes in frequencies cetrain values which appear to be almost equally spaced.
2. A reasonable assumption to make would be that for dogs whose birthdates were unknown, the birthdate was put in as the 1st of January of that particular year, or another common date, leading to sharp spikes in frequency for these birthdates.

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

```
ex<-filter(new,as.Date("1980-01-01") < as.Date(new$date) & as.Date(new$date)<as.Date(
"2012-06-26"))

hist(ex$date,"years",xlab="Birthdates",freq=TRUE)
```



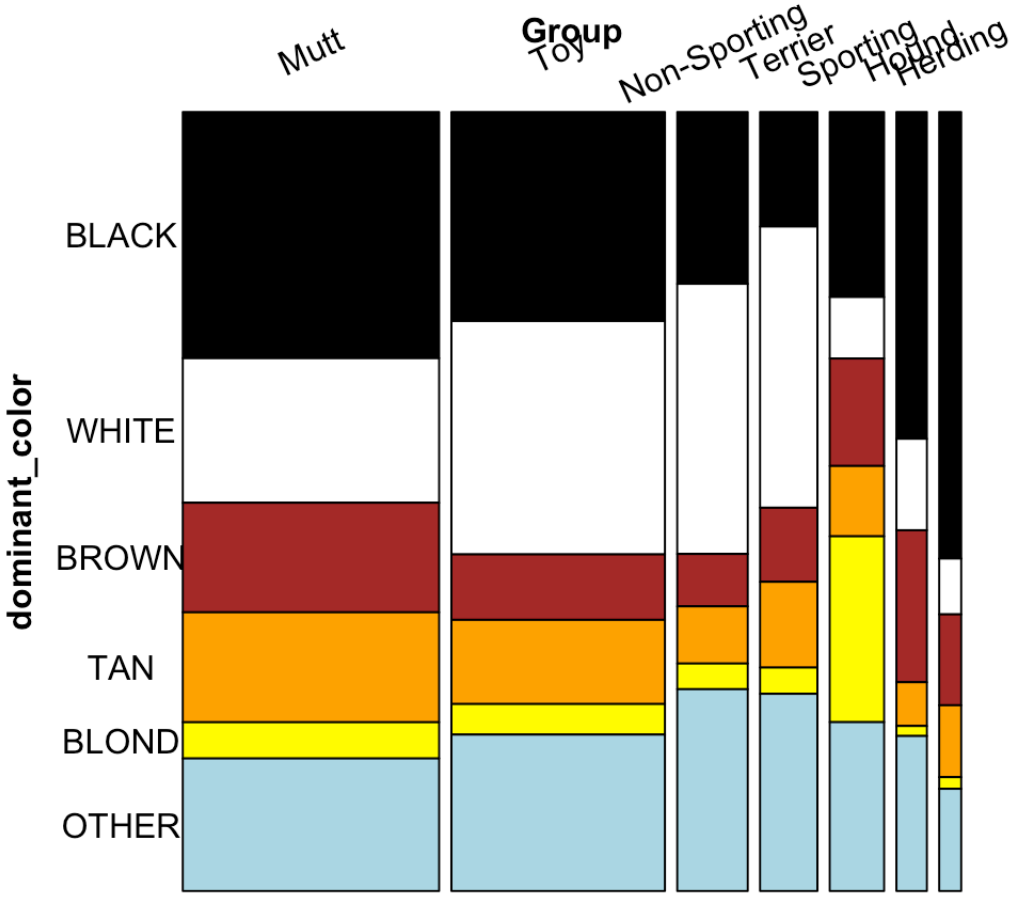
3. Mosaic plots

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.

```
library(vcd)
library(grid)
library(RColorBrewer)
fillcolors <- brewer.pal(6, "Set2")

df%>%group_by(dominant_color)%>%summarize(tot=n())%>%arrange(desc(tot))
```

```
## # A tibble: 20 x 2
##    dominant_color   tot
##    <chr>          <int>
##  1 BLACK          23578
##  2 WHITE          18621
##  3 BROWN           9181
##  4 TAN             8942
##  5 BLOND           4241
##  6 GRAY            2777
##  7 BRINDLE         2627
##  8 RUST            2263
##  9 BLUE            1378
## 10 FAWN            1367
## 11 RED             1338
## 12 CREAM           1186
## 13 GOLD            1015
## 14 <NA>             771
## 15 ORANGE           705
## 16 CHOCOLATE        538
## 17 APRICOT          468
## 18 SILVER           353
## 19 BLUE MERLE       103
## 20 CHARCOAL          90
```

```
mos<-mutate(df,dominant_color=ifelse((dominant_color=="BLACK"|dominant_color=="WHITE"
|dominant_color=="BROWN"|dominant_color=="TAN"|dominant_color=="BLOND"),dominant_colo
r,"OTHER"))

mos$Group<-factor(mos$Group, levels =c("Mutt","Toy","Non-Sporting","Terrier","Sportin
g","Hound","Herding"))

mos$dominant_color<-factor(mos$dominant_color, levels =c("BLACK","WHITE","BROWN","TAN
","BLOND","OTHER"))
myColors <- c("black", "white", "brown","orange","yellow","lightblue")
vcd::mosaic(dominant_color ~Group,mos,direction=c("v"),gp = gpar(fill = myColors),lab
eling = labeling_border(rot_labels = c(25, 0, 0, 0), offset_varnames = c(0,0,0,2),off
set_labels = c(0.5,0,0,0.5)))
```



1.Dominant color does appear to be dependent upon group.

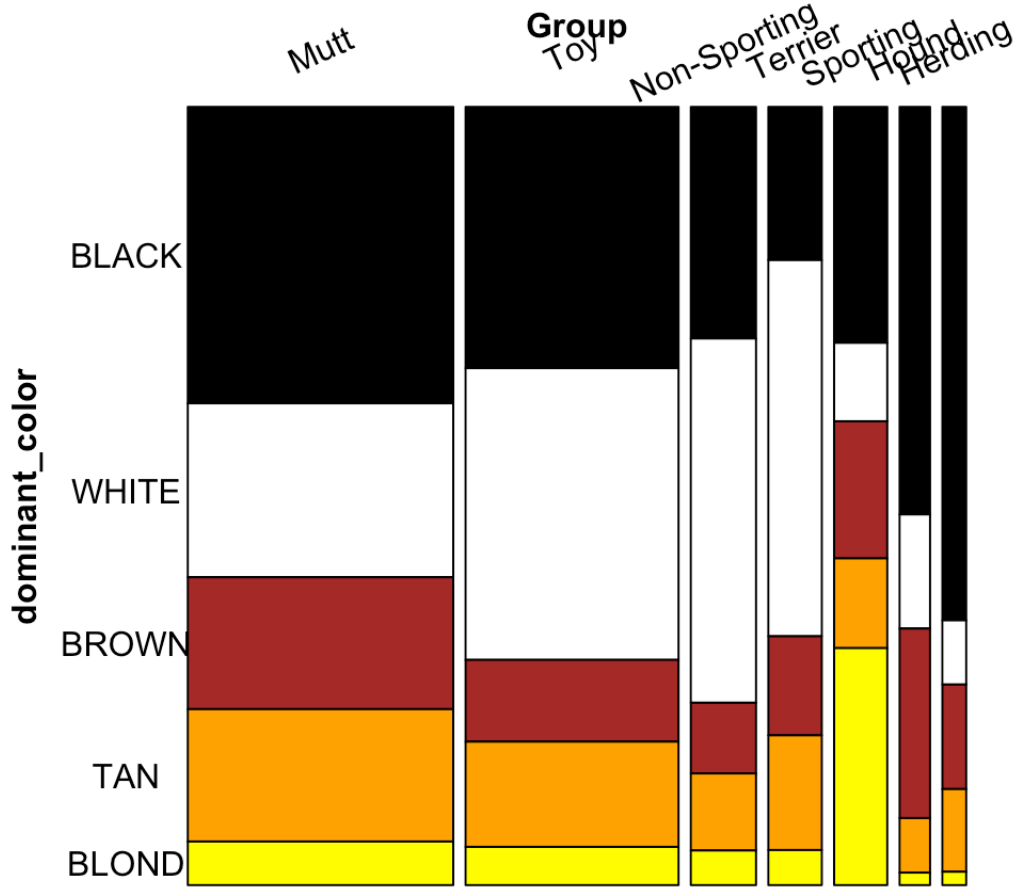
2.The most common dominant color by group varies. Non-sporting and Terrier Groups have more White dominant colored dogs compared to the other groups.

3.While Blond colored dogs are usually the least common amongst the top 5 colors, the Group of Sporting Dogs has an uncommonly high proportion of Blond dogs.

4.Black is the dominant color in most categories

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?

```
myColors <- c("black", "white", "brown","orange","yellow")
fillcolors <- brewer.pal(5, "Set2")
mod<-filter(mos,dominant_color!="OTHER")
mod$dominant_color<-factor(mod$dominant_color, levels =c("BLACK","WHITE","BROWN","TAN
","BLOND"))
vcd::mosaic(dominant_color ~Group,mod,direction=c("v"),gp = gpar(fill = myColors),lab
eling = labeling_border(rot_labels = c(25, 0, 0, 0), offset_varnames = c(0,0,0,2),off
set_labels = c(0.4,0,0,0.5)))
```



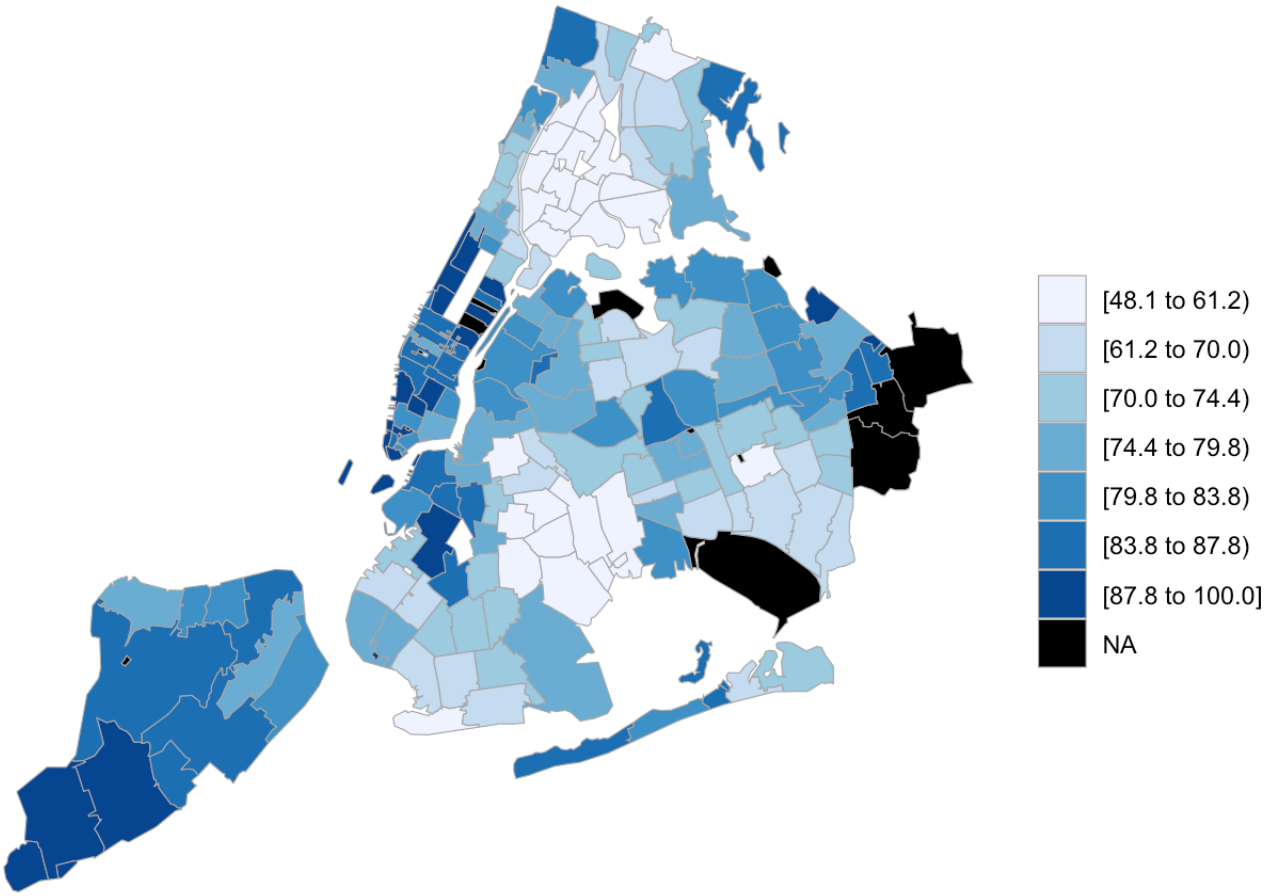
1.No the results do not change when the ‘Others’ category is removed.

2.If there are values in the dataframe which do not occur as frequently as certain values, then it would be safe to group them in the ‘Others’ category and display them. However, if the ‘Others’ category contains values that occur frequently, then its size becomes so large, that it doesn’t offer any additional insights into date and hence can be dropped.

4. Maps

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

```
library(choroplethrZip)

heat<-df%>%group_by(zip_code)%>%filter(spayed_or_neutered=="Yes")%>%summarize(n=n())


eg<-df%>%select(zip_code)%>%group_by(zip_code)%>%summarize(tot=n())

final<-left_join(x = eg, y = heat)
final<-final%>%mutate(per=(n/tot)*100)

names(final)[1]<-"region"
final$tot<-NULL
final$n<-NULL
names(final)[2]<-"value"
data(df_pop_zip)
final$region <- as.character(final$region)
nyc_fips = c(36005, 36047, 36061, 36081, 36085)
zip_choropleth(final,county_zoom=nyc_fips)
```



1.Some zipcodes on the peripheries of Queens seem to have missing values

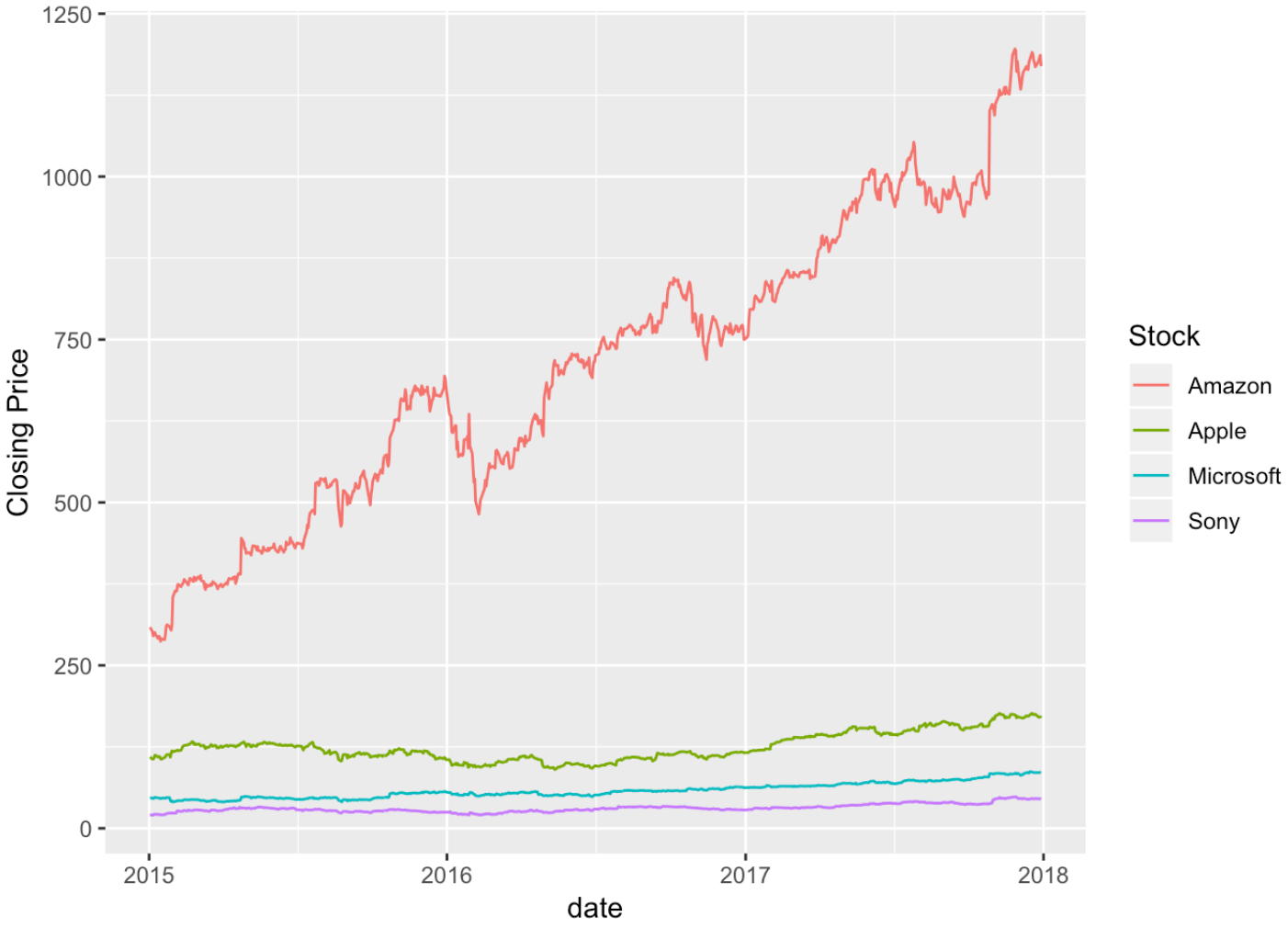
2.The Bronx contains the least percentage of Splayed or Neutered Dogs.

3.Staten Island and Manhattan, specifically Southern Staten Island and the Upper West Side and Downtown Manhattan have some of the highest percentages of splayed or neutered dogs.

5. Time Series

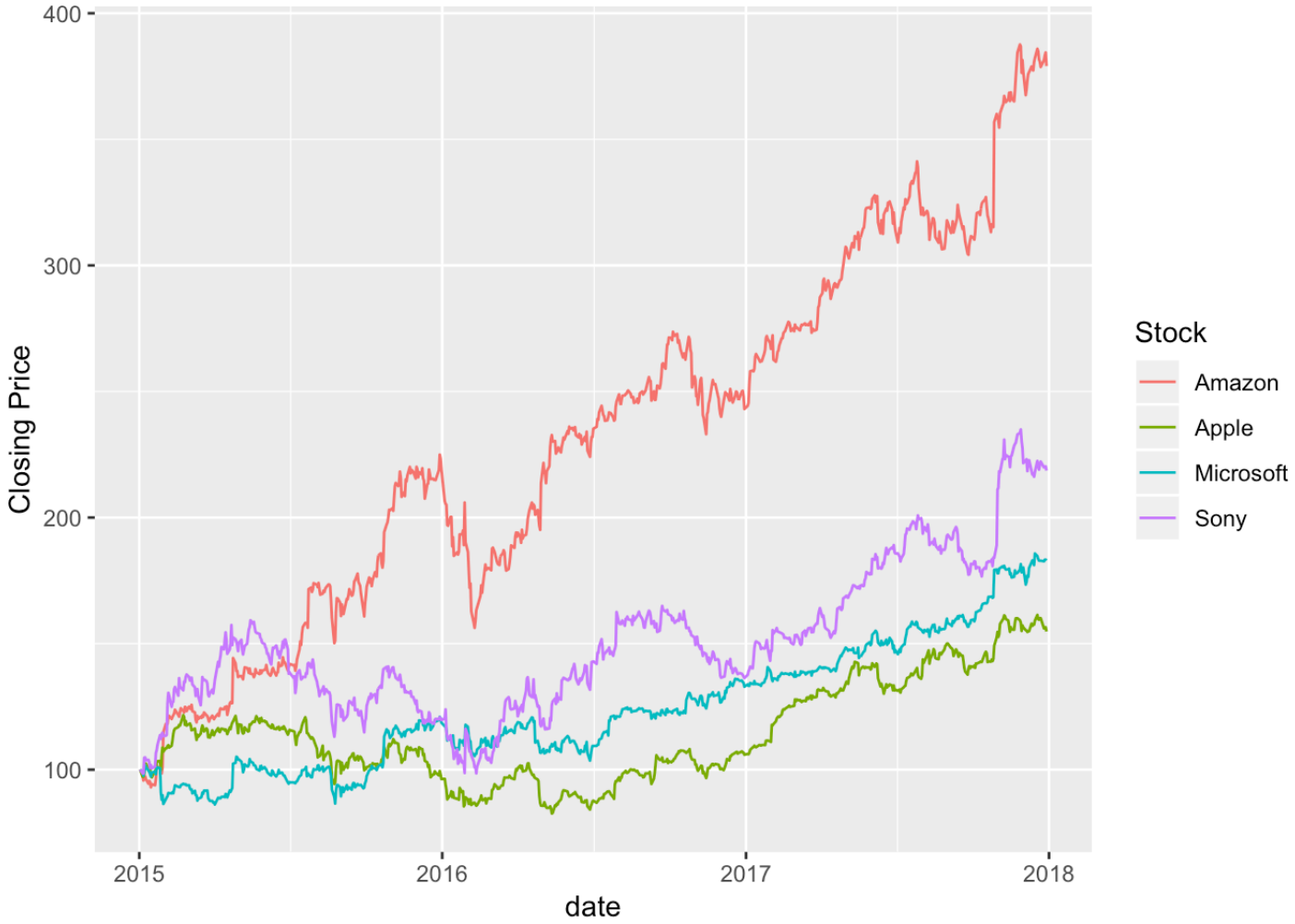
a. Use the `tidyquant` package to collect information on four tech stocks of your choosing. Create a multiple line chart of the closing prices of the four stocks on the same graph, showing each stock in a different color.

```
library(tidyquant)

AAPL <- tq_get("AAPL", get = "stock.prices", from = "2015-01-01", to = "2017-12-31")
AMZN <- tq_get("AMZN", get = "stock.prices", from = "2015-01-01", to = "2017-12-31")
MSFT<- tq_get("MSFT", get = "stock.prices", from = "2015-01-01", to = "2017-12-31")
SNE <- tq_get("SNE", get = "stock.prices", from = "2015-01-01", to = "2017-12-31")

apl<-AAPL%>%select(date,close)
amz<-AMZN%>%select(date,close)
mic<-MSFT%>%select(date,close)
sony<-SNE%>%select(date,close)
names(apl)[2]<-"Apple"
names(amz)[2]<-"Amazon"
names(mic)[2]<-"Microsoft"
names(sony)[2]<-"Sony"
t1<-left_join(apl,amz)
t2<-left_join(t1,mic)
t3<-left_join(t2,sony)
t4<-t3%>%gather(key = 'Stock', value = 'Close', Apple:Sony)
ggplot(t4,aes(x=date,y=Close,color=Stock))+geom_line()+ylab("Closing Price")
```



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)?

```
apl<-AAPL%>%select(date,close)
amz<-AMZN%>%select(date,close)
mic<-MSFT%>%select(date,close)
sony<-SNE%>%select(date,close)
apl<-apl%>%mutate(close=(close/apl$close[1])*100)
amz<-amz%>%mutate(close=(close/amz$close[1])*100)
mic<-mic%>%mutate(close=(close/mic$close[1])*100)
sony<-sony%>%mutate(close=(close/sony$close[1])*100)
names(apl)[2]<-"Apple"
names(amz)[2]<-"Amazon"
names(mic)[2]<-"Microsoft"
names(sony)[2]<-"Sony"
u1<-left_join(apl,amz)
u2<-left_join(u1,mic)
u3<-left_join(u2,sony)
u4<-u3%>%gather(key = 'Stock', value = 'Close', Apple:Sony)
ggplot(u4,aes(x=date,y=Close,color=Stock))+geom_line()+ylab("Closing Price")
```



1.After transforming the closing prices data, it can be observed that while the Amazon stock follows the same general trend, Sony outperforms the other two stocks over the same time period.

2.While the first graph makes it look like Sony and Microsoft closing prices haven't changed dramatically over 3 years, the tranformed graphs shows that they have made considerable gains when compared to their baselines.

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…)

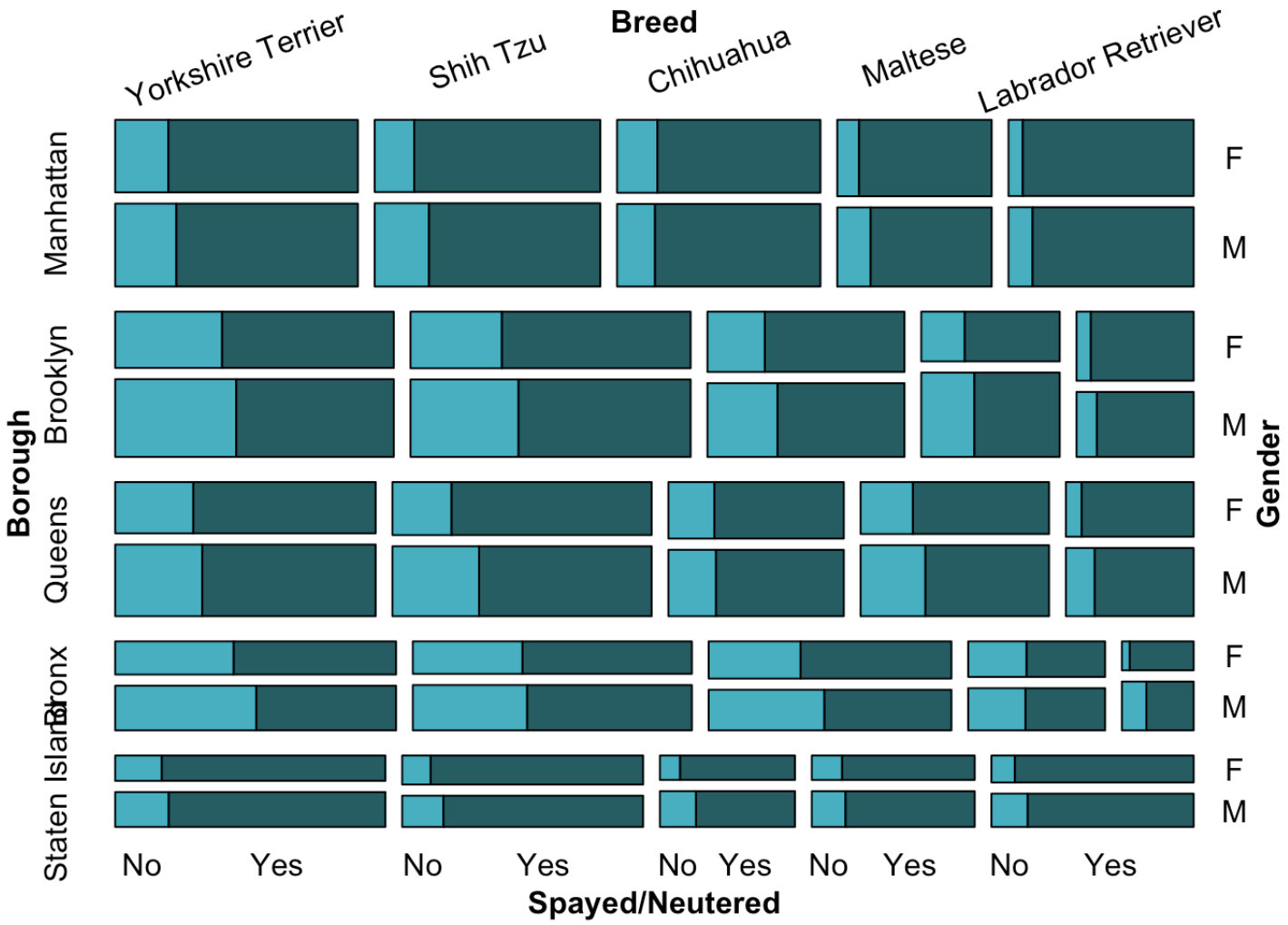
A construction company that wishes to develop a policy around pets being allowed into buildings it has built in specific boroughs.

b. What is the main point you hope someone will take away from the graph?

1. The main point is that 4 out of 5 most common breeds in NYC are small dogs, which is convenient in a city with small apartments. Additionally, small dogs are easier to take on subways.
2. Additionally, the mosaic plot gives the proportion of each breed according to borough and its distribution according to whether or not they are spayed or neutered.

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.

```
bred<-df%>%filter(breed=="Labrador Retriever"|breed=="Yorkshire Terrier"|breed=="Shih
Tzu"|breed=="Chihuahua"|breed=="Maltese")



bred$breed<-factor(bred$breed, levels =c("Yorkshire Terrier","Shih Tzu","Chihuahua","
Maltese","Labrador Retriever"))


bred$borough<-factor(bred$borough, levels =c("Manhattan","Brooklyn","Queens","Bronx",
"Staten Island"))

fillcolor <- brewer.pal(2, "Set2")
vcd::mosaic(spayed_or_neutered ~borough+breed+gender,bred,gp = gpar(fill = c("#48b0c2
","#275e66")),labeling = labeling_border(set_varnames=c(borough="Borough",breed="Bree
d",gender="Gender",spayed_or_neutered="Spayed/Neutered")
,rot_labels = c(20, 0, 0, 90),offset_labels = c(0.65,0,0,0.6)))
```


Breed
Yorkshire Terrier
Shih Tzu
Chihuahua
Maltese
Labrador Retriever
Borough
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Gender
F
M
No
Yes
Spayed/Neutered