HW03

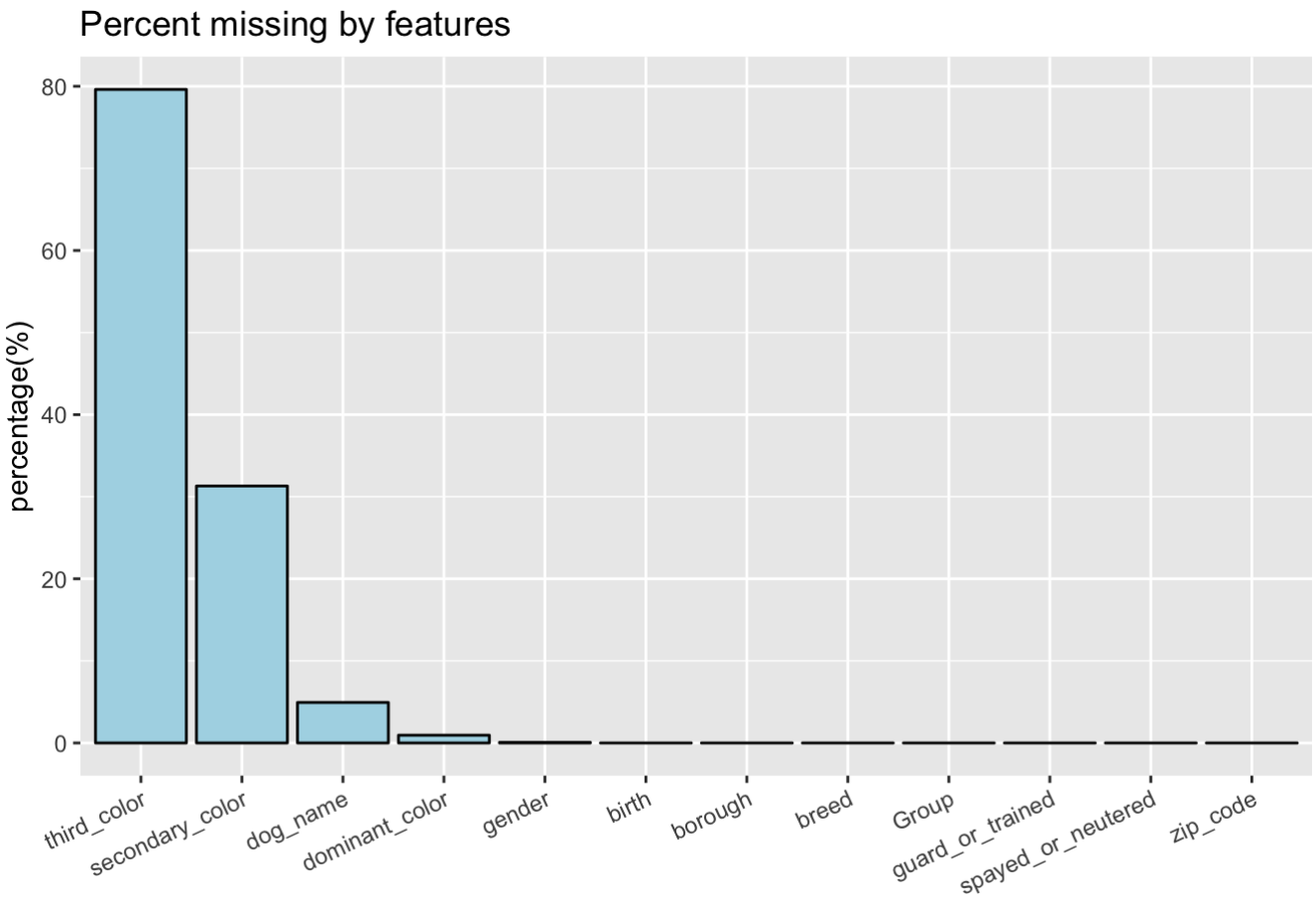
Yawen Han (yh3069)

10/26/2018

Problem 1 - Missing Data

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

```
library(tidyverse)
# load data from csv file
dogs <- read_csv("/Users/yawenhan/Desktop/Autumn2018/5702 EDAV/R Code/HW03/NYCdogs.csv")
# transform missing value as "NA"
dogs[dogs == 'n/a'] <- NA
# find out the missing count and percentage
colNa <- data.frame(colSums(is.na(dogs)) %>% sort(decreasing = TRUE))
colnames(colNa) <- c("countNa")
colNa$features <- rownames(colNa)
colNa$percentNa <- colNa$countNa/length(dogs$dog_name)*100
# plot the barchart of percent missing by features with a decreasing order
ggplot(colNa,aes(x=fct_reorder(features,percentNa,.desc=TRUE),y=percentNa))+geom_bar(col
or="black",fill="lightblue",stat="identity")+theme(axis.text.x=element_text(angle=25,hju
st=1))+ylab("percentage(%)")+ggtitle("Percent missing by features")+xlab("")
```

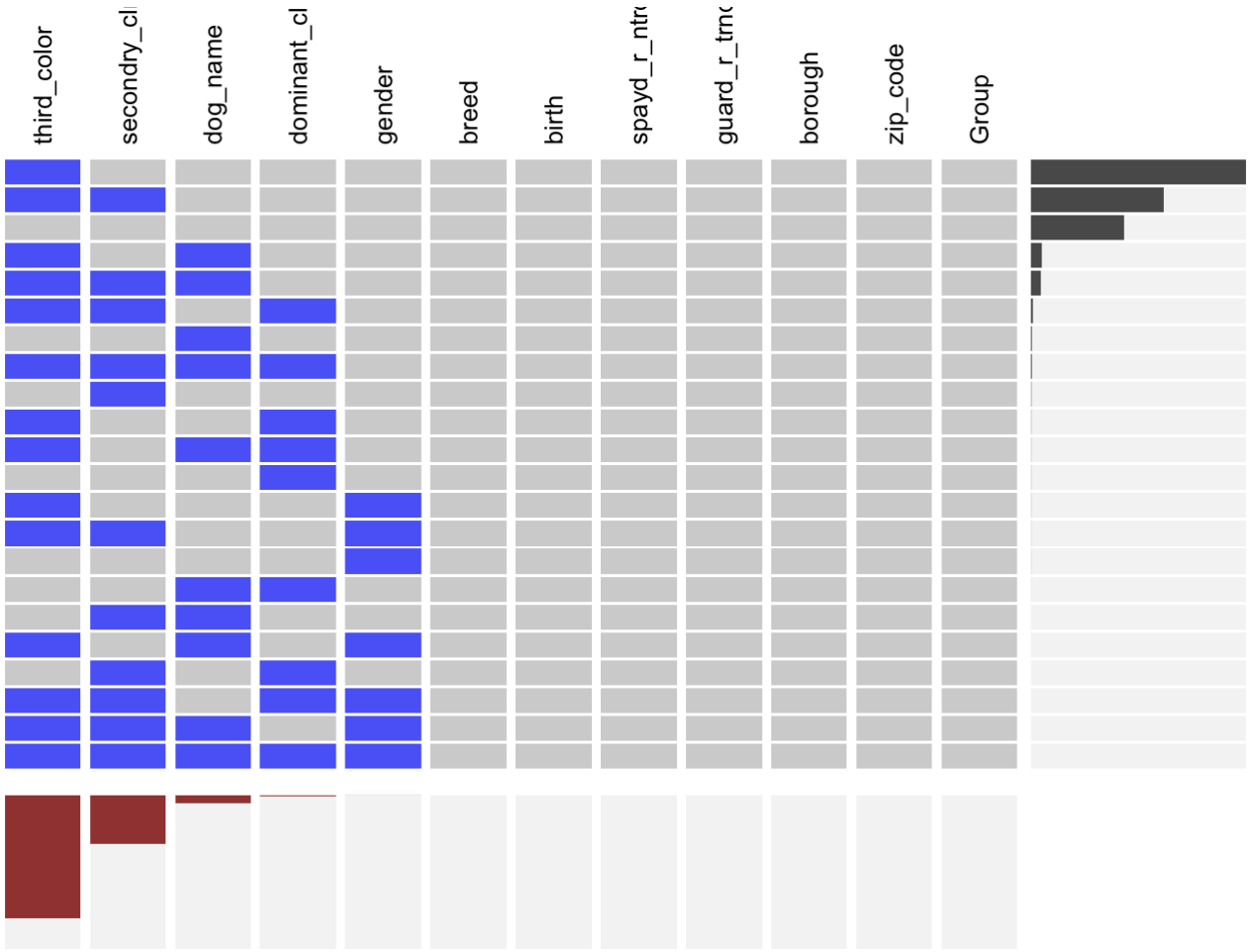


Before plot the bar chart of missing percentage, it is observed that the missing value is represented as “n/a” in the provided dataset. For easier missing value detection, all missing values are transformed as “NA”; then the “colNa” dataframe is created, which consists three columns: “features”, “countNa” and “percentNa”. Finally, plot the bar chart of missing percentage by features with decreasing order, to give a better understanding of the missing value distribution.

In the bar chart above, “third_color” contains the most missing value. Nearly 80% of the data has a missing value on “thrid_color”. The second is “secondary_color” with around 30% missing, then other features with less than 10% missing.

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

```
library(extracat)
visna(dogs, sort = "b")
```



The extracat:visna plot above allows us to visualize missing patterns of “NYCdogs” data. The columns represent the 12 variables and the rows the missing patterns. The cells for the variables with missing values in a pattern are drawn in blue. The variables and patterns have been ordered by numbers of missings on both rows and columns (sort = “b”). The bars beneath the columns show the proportions of missings by variable and the bars on the right show the relative frequencies of patterns.

It is observed from the plot that:
(1)The bars beneath the columns show the proportions of missings by variable, which is the same as the bar chart in Problem 1-a. Top three missing variables are “third_color”, “secondary_color” and “dog name”.
(2)The bars on the right show the relative frequencies of patterns. The top three patterns are missing “third_color”, missing “third_color” & “secondary_color”, and no missing values.

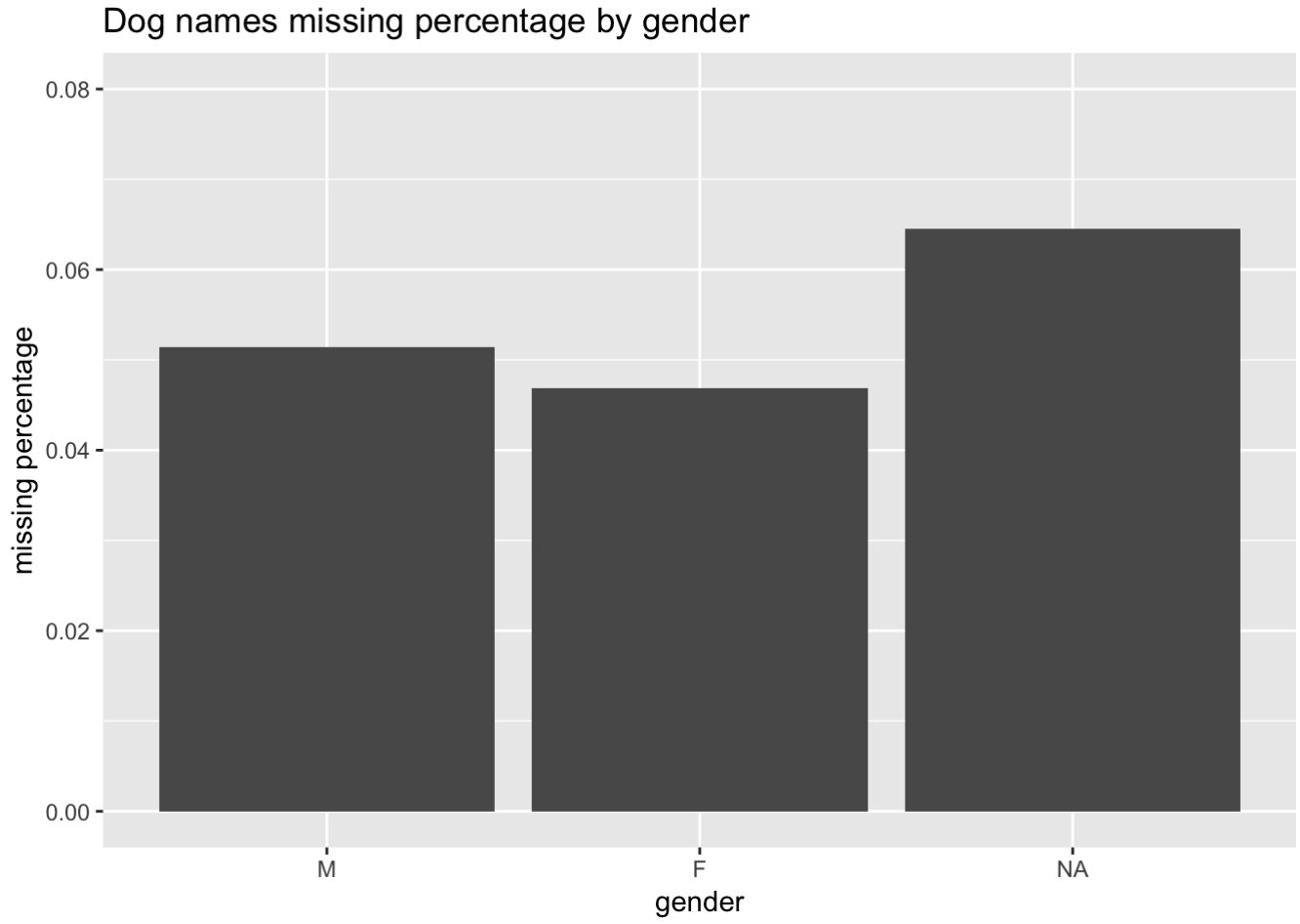
(c)Do dog_name missing patterns appear to be associated with the value of gender,Group or borough?

1. First, explore the dog_name missing patterns with the missing value of gender, Group or borough:

From the extracat:visna plot in Problem 1-b, it can be concluded:
(1)There is no missing value for “Group” and “borough”, thus no “dog_name” missing patterns appear to be associated with the missing value of “Group” and “borough”.
(2)There are three “dog_name” missing patterns appear to be associated with the value of the missing value of “gender”, including
* a. “third_color”, “dog_name” and “gender”;
* b. “third_color”, “secondary_color”, “dog_name” and “gender”;
* c. “third_color”, “secondary_color”, “dog_name”, “dominate_color” and “gender”;

2. Then, exploring the dog_name missing patterns with different level values of gender, Group or borough:

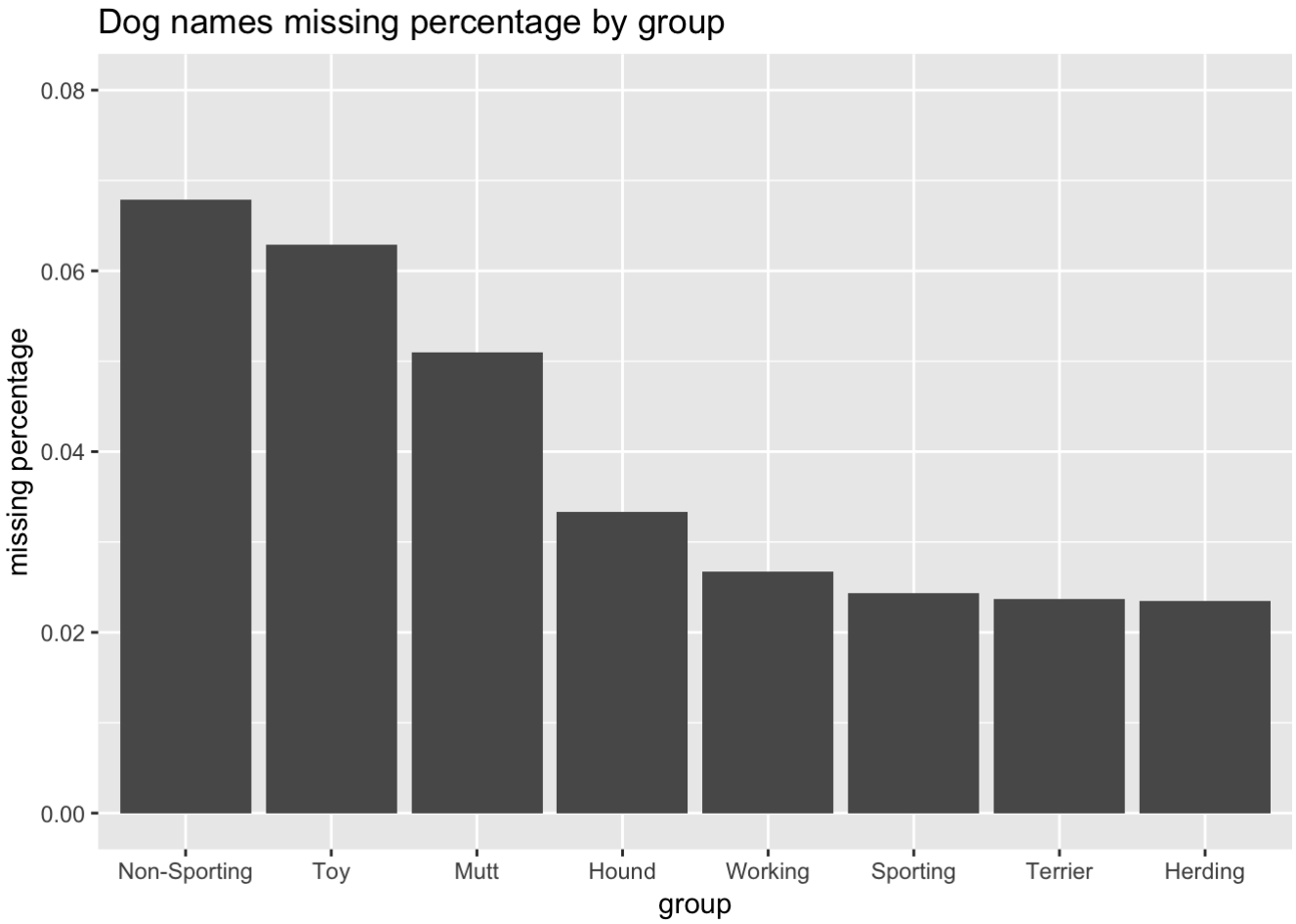
```
missing_gender<-dogs %>% group_by(gender) %>% summarise(total=sum(is.na(dog_name)),n=n
())
missing_gender$percentage<-missing_gender$total/missing_gender$n
ggplot(missing_gender)+geom_bar(aes(x=fct_reorder(gender,percentage,.desc=TRUE),y=percen
tage),stat='identity')+ggtitle("Dog names missing percentage by gender")+xlab("gender")+
ylab("missing percentage")+ylim(c(0,0.08))
```



Plot the bar chart above to explore the dog_name missing patterns with the value of gender. As the difference of missing percentage between “F” and “M” is not very significant, the association with two level values seems weak. However, if the gender has a missing value, the dog_name has a higher percentage of missing, which re-

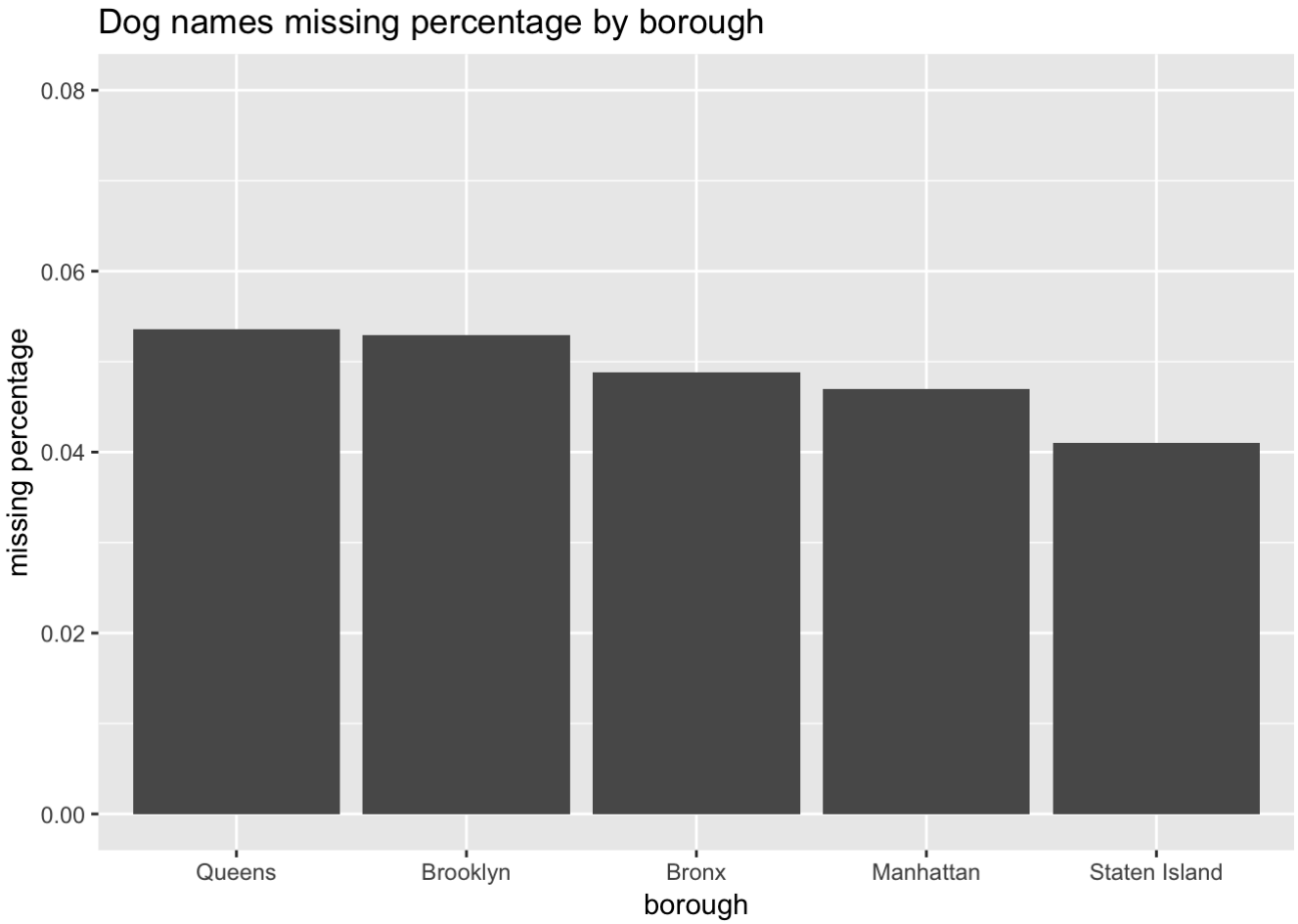
confirms the conclusion in part1.

```
missing_group<-dogs %>% group_by(Group) %>% summarise(total=sum(is.na(dog_name)),n=n())
missing_group$percentage<-missing_group$total/missing_group$n
ggplot(missing_group)+geom_bar(aes(x=fct_reorder(Group,percentage,.desc=TRUE),y=percentage),stat='identity')+ggtitle("Dog names missing percentage by group")+xlab("group")+ylab("missing percentage")+ylim(c(0,0.08))
```



Plot the bar chart above to explore the dog_name missing patterns with the value of the group. As the difference of missing percentage between different group is significant, the association between missing names and group is strong.

```
missing_borough<-dogs %>% group_by(borough) %>% summarise(total=sum(is.na(dog_name)),n=n())
missing_borough$percentage<-missing_borough$total/missing_borough$n
ggplot(missing_borough)+geom_bar(aes(x=fct_reorder(borough,percentage,.desc=TRUE),y=percentage),stat='identity')+ggtitle("Dog names missing percentage by borough")+xlab("borough")+ylab("missing percentage")+ylim(c(0,0.08))
```



Plot the bar chart above to explore the dog_name missing patterns with the value of the borough. As the difference of missing percentage between different borough is not very significant, the association between missing names and borough seems weak.

Problem 2 - Dates

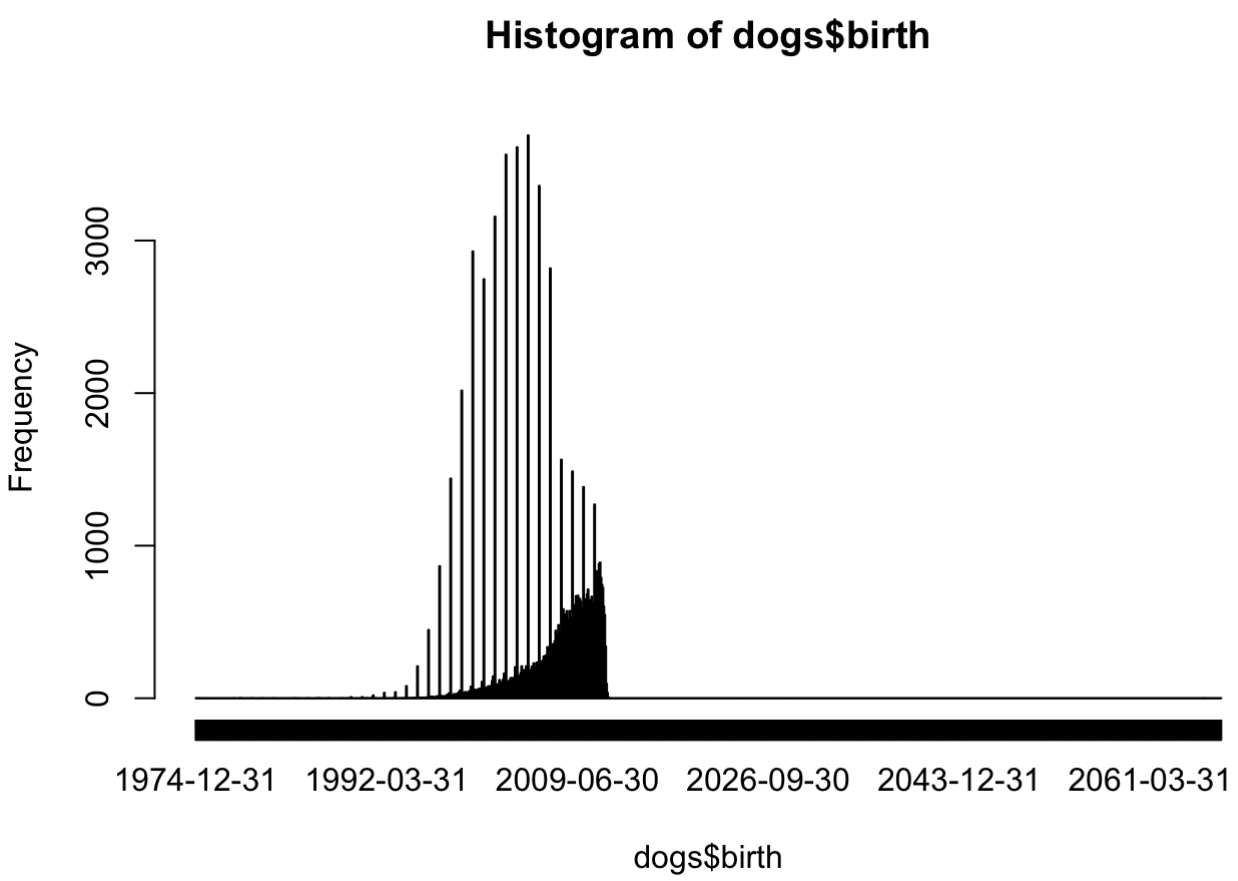
(a) Convert the birth column of the NYC dogs dataset to Date class (use “01” for the day since it’s not provided). Create a frequency histogram of birthdates with a one-month binwidth. (Hint: don’t forget about base R.) What do you observe? Provide a reasonable hypothesis for the prominent pattern in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

```
library(stringr)
dogs$month<-str_extract(dogs$birth, "[a-zA-Z]+")#extract birth "month" from the "birth"
dogs$year<-str_extract_all(dogs$birth,"\\d+")#extract birth "year"" from the "birth"
dogs$year<-str_pad(dogs$year, 2, pad = "0")
dogs$day<-rep("01",length(dogs$birth))#use “01” for the birth "day"" since it’s not prov
ided
dogs$birth<-str_c(dogs$month,dogs$day,dogs$year,sep="-") #combine "month", “day” and "ye
ar" together as the "birth" column
dogs$birth <- as.Date(dogs$birth,format = "%b-%d-%y")#Convert the birth column of the NY
C dogs dataset to Date class
class(dogs$birth)
```

```
## [1] "Date"
```

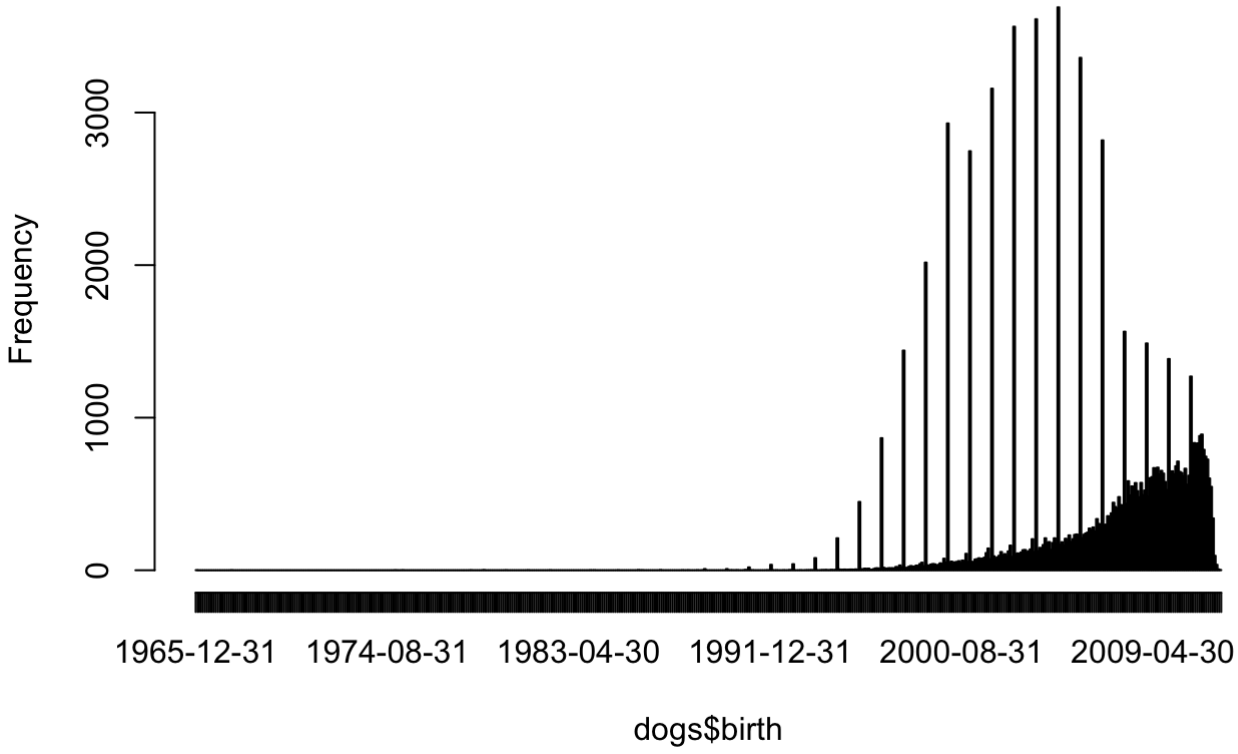
In the “NYCdogs” dataset, the variable “birth” owns inconsistent formats: some instance has a birth value “Jan-00” as “January 2000”, while some instance has a birth value “2-Jan” as “January 2002”. To uniform the date format, the code above extract birth “month” and “year” from the variable “birth”, and add a “day” column all with value “01”.Then combine “month”, “day” and “year” together to update the “birth” column, and convert the “birth” column of the NYC dogs dataset to Date class. The output above shows the class of “birth” is “Date”.

```
library(lubridate)
hist(dogs$birth, "months", freq = TRUE)
```



The histogram above is a frequency histogram of birthdates with a one-month binwidth. Since we should not have data after the current date, it’s obvious that there exist wrong data values. Considering the fact that the “as.Date()” can only transform the date after “1970-01-01”, some data before the year 1970 are transformed to be 100 years later. Therefore, these data values are corrected, then replot the frequency histogram of birthdates.

```
wrong_year<-year(dogs$birth[which(dogs$birth>as.Date("2018-01-01"), arr.ind=TRUE)])#find
 out the wrong year transformed by lubridate
year(dogs$birth[which(dogs$birth>as.Date("2018-01-01"), arr.ind=TRUE)])<-wrong_year-100#
correct the wrong years
hist(dogs$birth, "month", freq = TRUE)
```

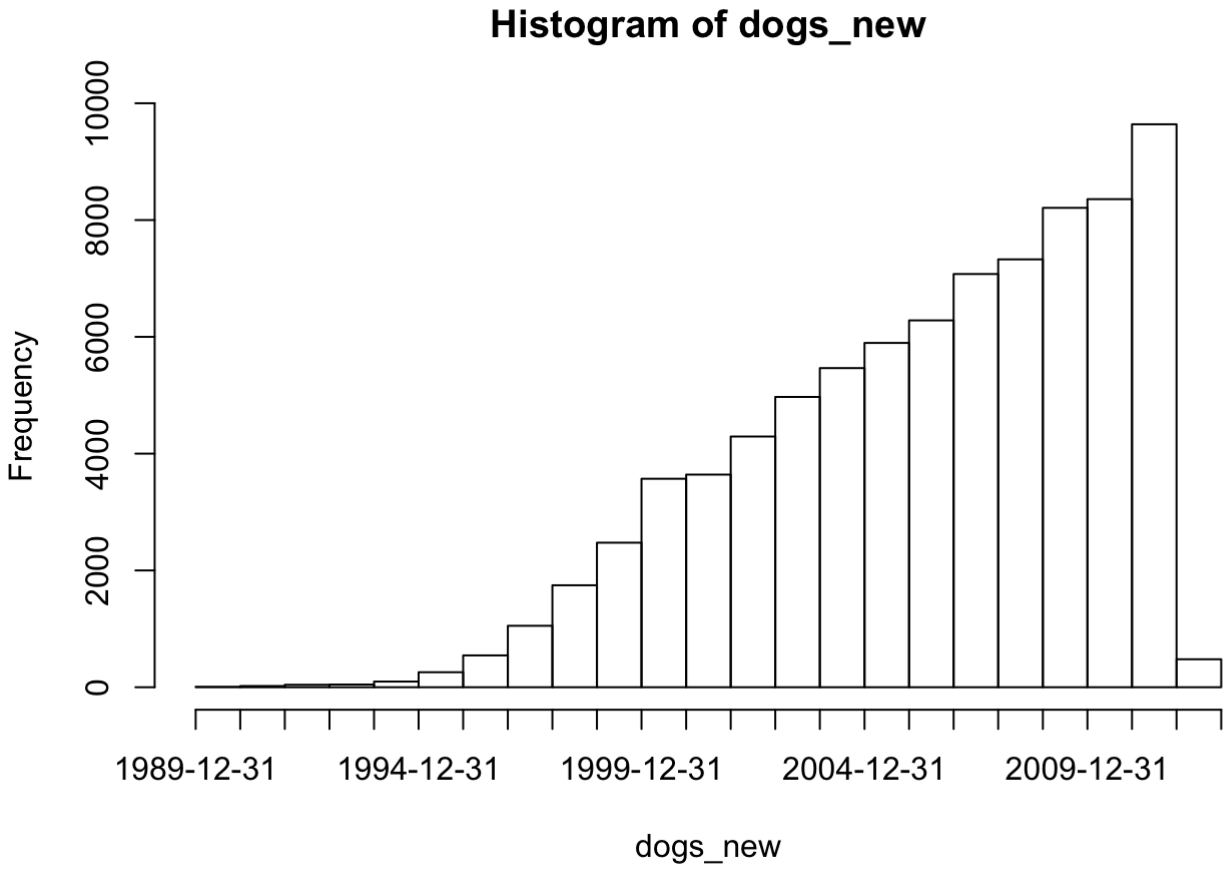


The histogram above is the updated frequency histogram of birthdates with a one-month binwidth. There are several patterns are observed from the histogram:
(1)Some dogs are born more than 20 years ago, which should already die in common sense;
(2)Two separate distributions appear in the histogram,
* a.the upper one like a “bell” shape with a very high frequency in “January” for all years; one possible reason may be that the owners choose to use “January” as a birthdate if they don’t know that;
* b.the lower one with an increasing trend with the year;

With a hypothesis that “the owners choose to use ‘January’ as a birthdate if they don’t know their dogs’ real birthdate”, plot the histogram with “months” as a binwidth is misleading. A more reasonable binwidth will be used in section b.

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

```
dogs_new<-dogs$birth[which(dogs$birth>as.Date("1990-01-01"), arr.ind=TRUE)]
hist(dogs_new, "years", freq = TRUE)
```

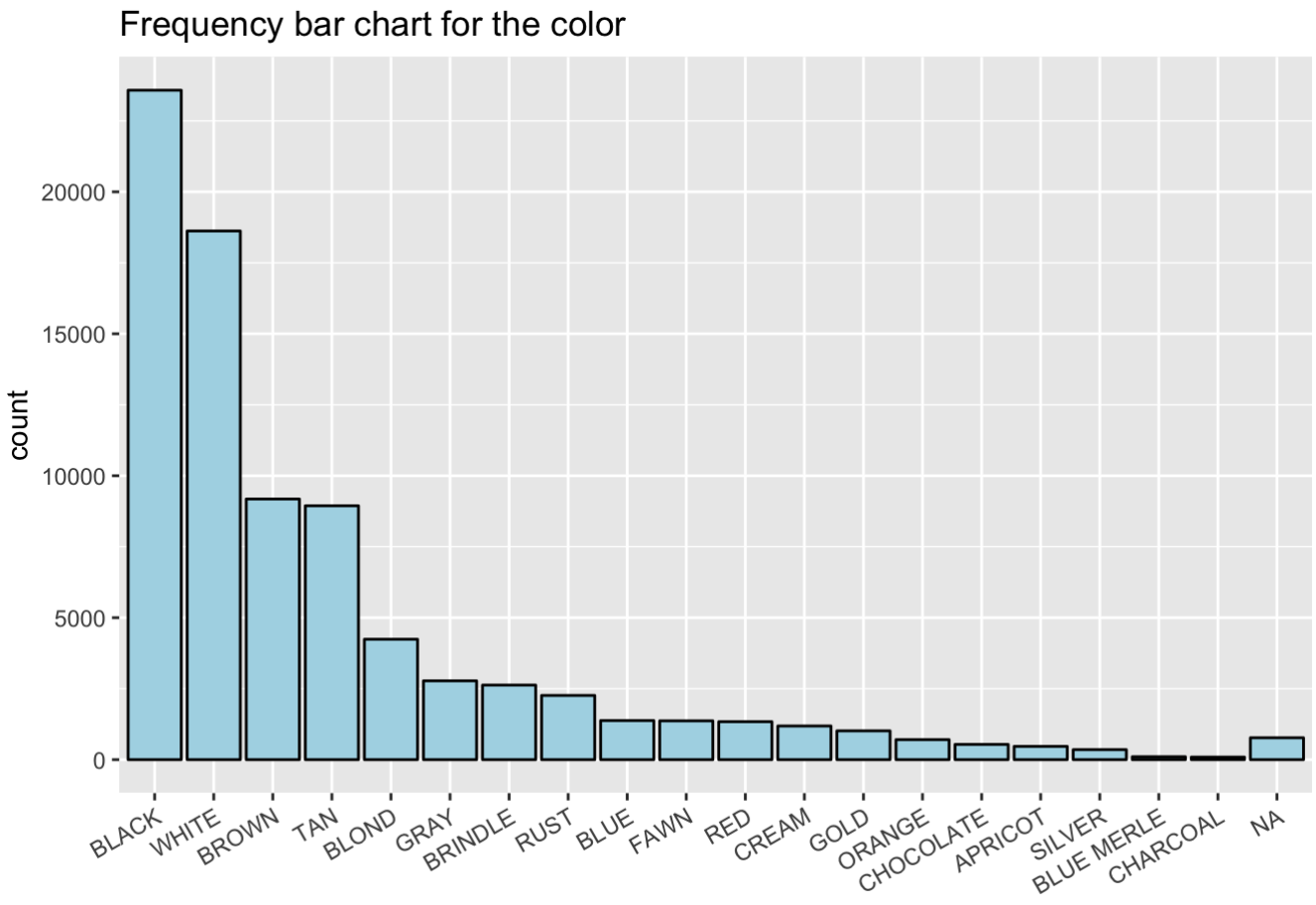


According to the observation in section a, we suppose that all the dogs born before 1990-01-01 are died and drop these data value. Then choose “year” as a new binwidth to eliminate the influence of the high “January” birthdate frequency. The new histogram of birthdates with a one-year binwidth is shown above. Actually, it shows an increasing pattern of dog number in NYC with the year.

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.

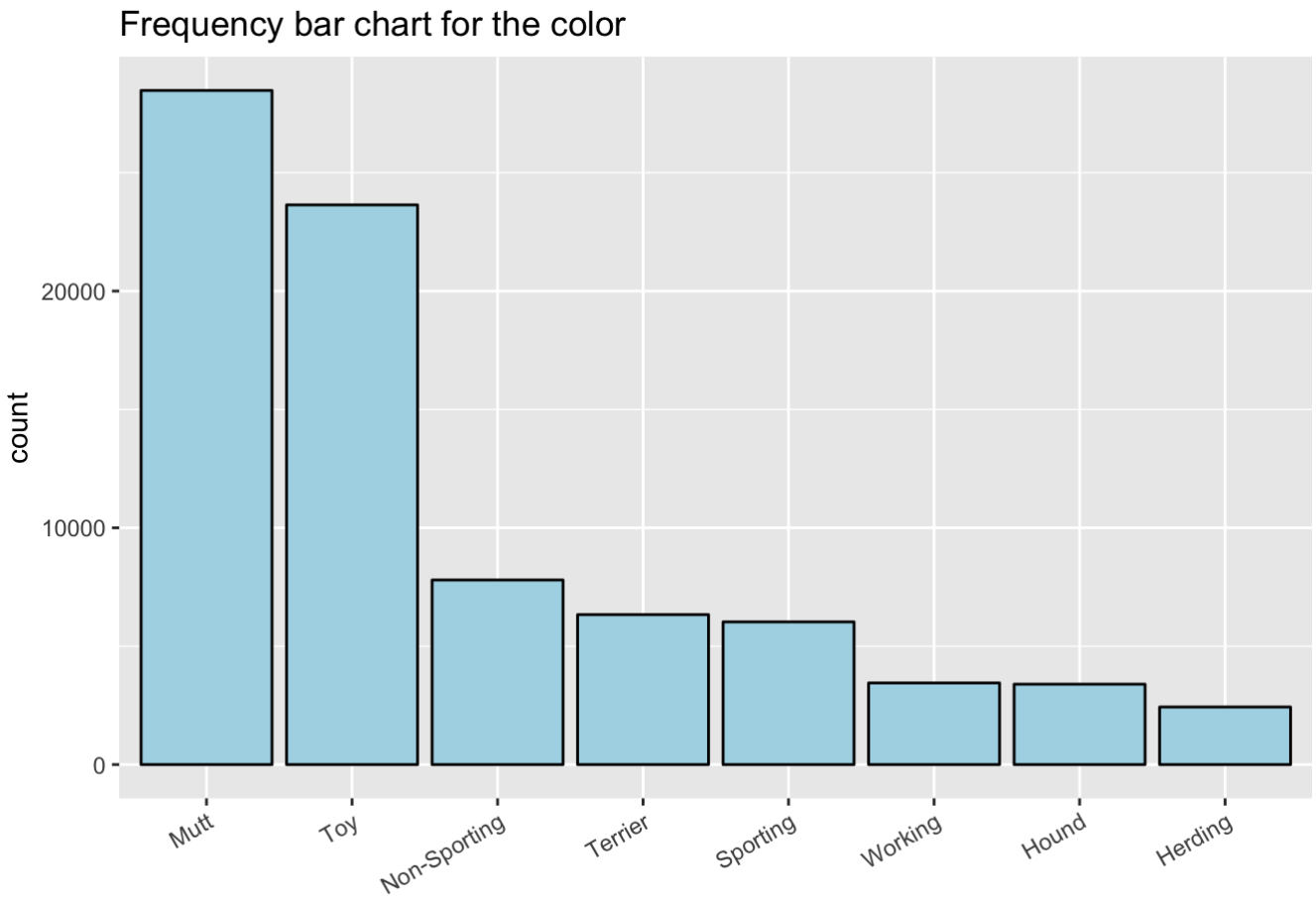
```
#factor(dogs$dominant_color)
ggplot(dogs, aes(x=reorder(dominant_color,dominant_color,function(x)-length(x))))+geom_b
ar(color="black",fill="lightblue")+ggtitle("Frequency bar chart for the color")+xlab("")
+theme(axis.text.x=element_text(angle=30,hjust=1))
```



According to the bar chart above, the top 5 dominant colors are “BLACK”, “WHITE”, “BROWN”, “TAN”, “BLOND”.

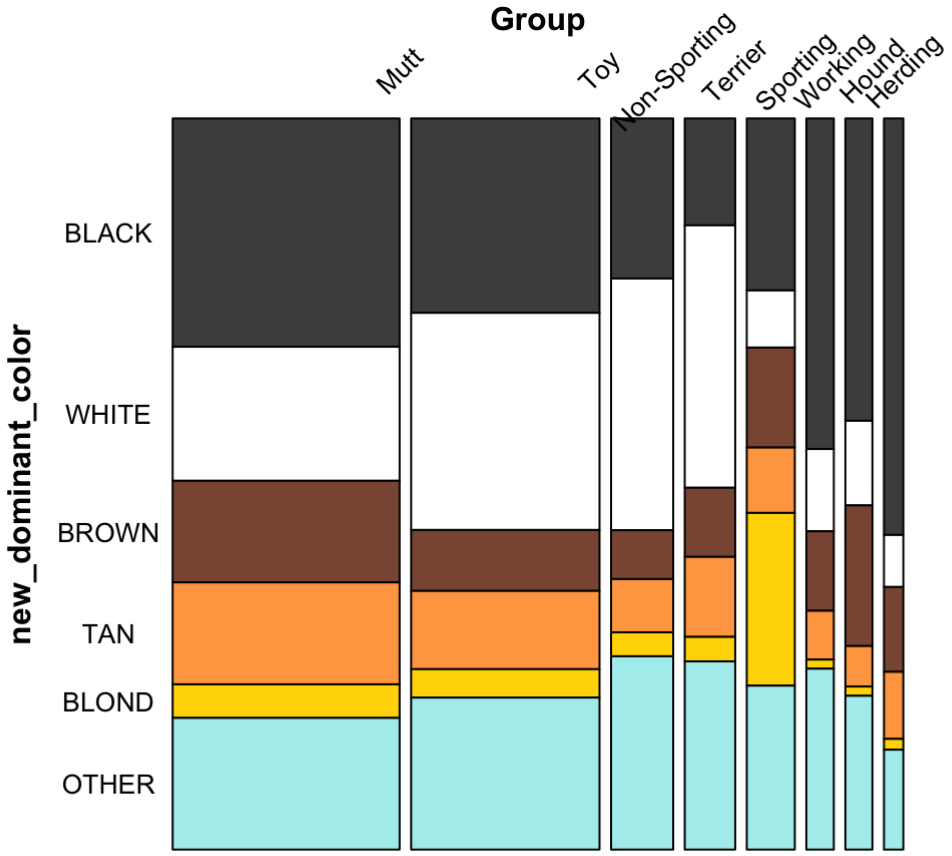
Then we try to find out the order of variable “Group”:

```
#factor(dogs$dominant_color)
ggplot(dogs, aes(x=reorder(Group,Group,function(x)-length(x))))+geom_bar(color="black",f
ill="lightblue")+ggtitle("Frequency bar chart for the color")+xlab("")+theme(axis.text.x
=element_text(angle=30,hjust=1))
```



According to the bar chart above, the order of variable “Group” is “Mutt”, “Toy”, “Non-Sporting”, “Terrier”, “Sporting”, “Working”, “Hound”, and “Herding”.

```
max_color<-c("BLACK","WHITE","BROWN","TAN","BLOND") #top 5 dominant colors
dogs$new_dominant_color<-dogs$dominant_color
dogs$new_dominant_color[which(!dogs$dominant_color %in% max_color)]<-"OTHER" #group the
 rest into an “OTHER” category
library(vcd)
library(grid) # needed for gpar
#library(RColorBrewer)
fillcolors <- c("gray30","white","lightsalmon4","tan1","gold","paleturquoise2") # give e
ach label corresponding colors
dogs$new_dominant_color<-factor(dogs$new_dominant_color,levels= c( "BLACK", "WHITE", "BR
OWN", "TAN", "BLOND", "OTHER"))
dogs$Group<-factor(dogs$Group,levels=c("Mutt","Toy","Non-Sporting","Terrier","Sporting",
"Working","Hound","Herding"))
vcd::mosaic( new_dominant_color ~ Group, dogs, gp = gpar(fill = fillcolors), direction =
 c("v", "h"),tl_labels = c(TRUE, TRUE),labeling = labeling_border(gp_labels = gpar(fonts
ize = 10),gp_varnames = gpar(fontsize = 12,fontface = 2), rot_labels = c(45, 0, 0, 0), o
ffset_varnames = c(0.6,0,0,2), offset_labels=c(0.5,0,0,1), pos_labels = c("right", "cent
er", "left", "center")))
```

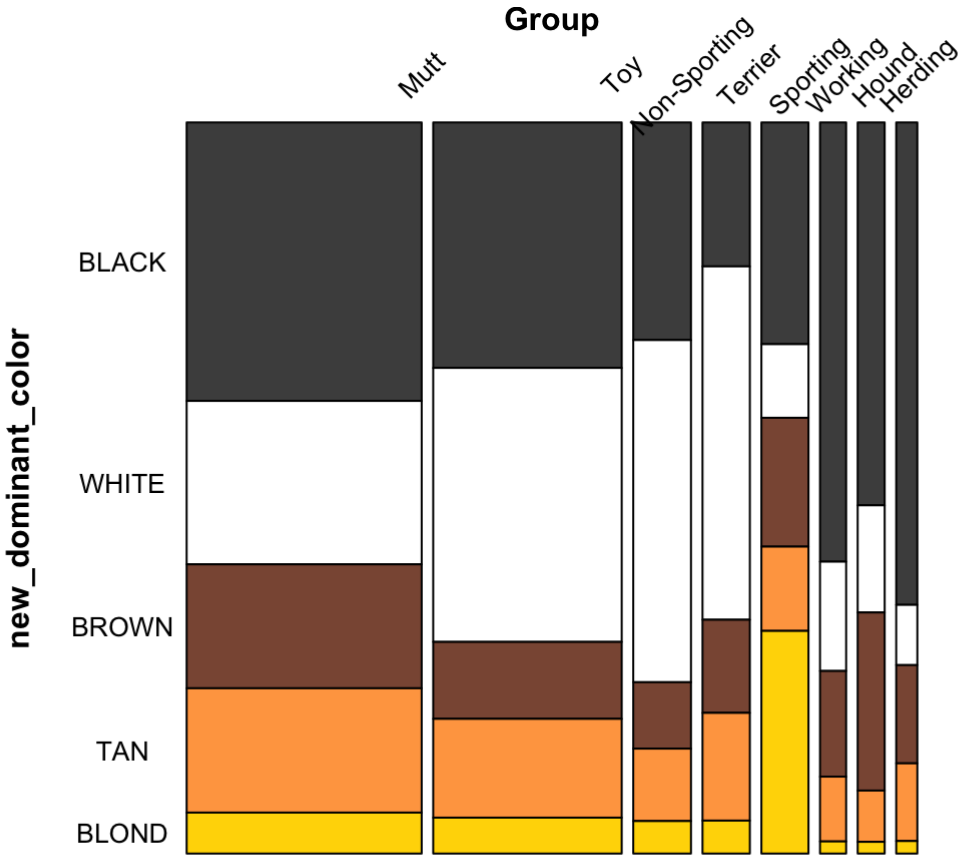


The graph above is a mosaic plot of “dominant_color” and “Group”, in which “dominant_color” is the dependent variable with top 5 dominant colors the rest as an “OTHER” category. The “Group” is also plotted in decreasing order from left to right.

In the plot above, it’s easy to observe the various color proportions for the different group. Thus, there is an associated relationship between two variables.

(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_no_other<-dogs[-c(which(!dogs$dominant_color %in% max_color)),]
fillcolors2 <- c("gray30","white","lightsalmon4","tan1","gold")
dogs_no_other$new_dominant_color<-factor(dogs_no_other$new_dominant_color,levels= c( "BL
ACK", "WHITE", "BROWN", "TAN", "BLOND"))
dogs_no_other$Group<-factor(dogs_no_other$Group,levels=c("Mutt","Toy","Non-Sporting","Te
rrier","Sporting","Working","Hound","Herding"))
vcd::mosaic( new_dominant_color ~ Group, dogs_no_other, gp = gpar(fill = fillcolors2), t
l_labels = c(TRUE, TRUE),direction = c("v", "h"), labeling = labeling_border(gp_labels =
 gpar(fontsize = 10),gp_varnames = gpar(fontsize = 12,fontface = 2), rot_labels = c(45,
0, 0, 0), offset_varnames = c(0.7,0,0,2.4), offset_labels=c(0.5,0,0,1), pos_labels = c(
"right", "center", "left", "center")))
```



Redraw the mosaic plot between “new_dominant_color” and “Group” by dropping the “OTHER” category from the “new_dominant_color” variable. As each proportion is enlarged without “OTHER” category, it’s also easy to observe the various color proportions for the different group. Thus, there is an associated relationship between two variables when considering the top 5 dominant colors and the dogs group.

When dropping the “OTHER” category, the proportion of each color = (1-proportion of OTHER)*relative ratio. Thus, the relative ratio between the top five colors is enlarged for easier comparison.

When deciding whether it’s necessary or not to include an “OTHER” category, I think we need to consider the ultimate goal of the analysis:
(1)If we only need to know the relationship between the most popular dominant color and the dogs group, the excluded mosaic plot is a good choice by eliminating the influence from “OTHER” category. For example, the “OTHER” category has a very large proportion in all groups, making it’s hard to observe the patterns;
(2)If we need to consider the overall relationship between two variables, only the top 5 dominant colors might not representative enough. For example, if only one group has a very high proportion in “OTHER” category while the rest groups have a low one, the drop of “OTHER” will affect the final conclusions significantly.

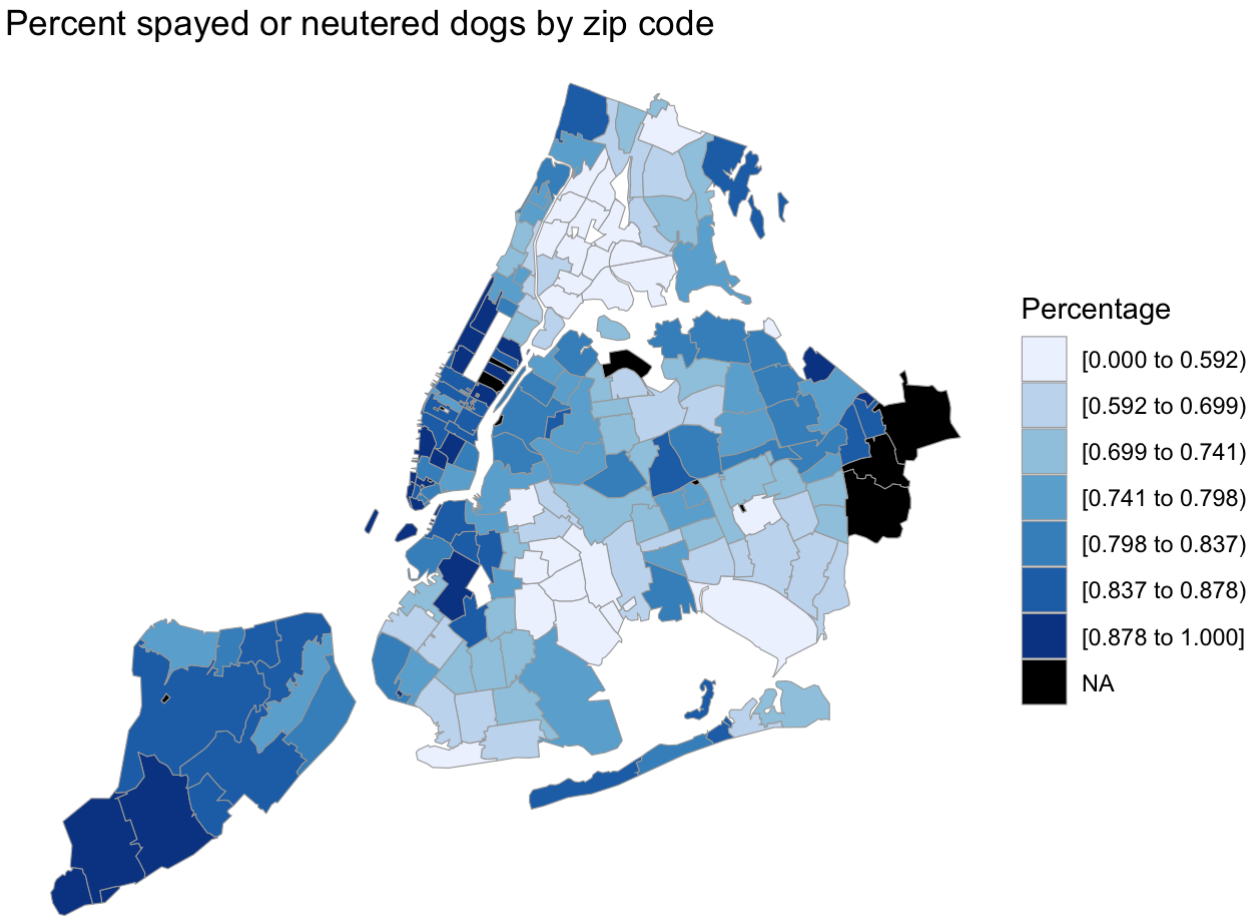
4. Maps

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

```
# group the gods data to fulfill the zip_choropleth requirement
dogs_table<-dogs %>% select(zip_code,spayed_or_neutered)
dogs_table$spayed_or_neutered<-ifelse(dogs_table$spayed_or_neutered=='Yes',1,0)
dogs_summary<-dogs_table%>% group_by(zip_code) %>%summarise(total=sum(as.numeric(spayed_
or_neutered)),n=n())
# calculate percentage
dogs_summary$value<-dogs_summary$total/dogs_summary$n
dogs_summary<-rename(dogs_summary,"region"="zip_code")
dogs_summary <- select(dogs_summary, -c("total","n"))
# change class
dogs_summary$region<-as.character(dogs_summary$region)
dogs_summary$value<-as.numeric(dogs_summary$value)
```

```
library(devtools)
library(choroplethr)
library(choroplethrZip)
nyc_fips = c("36005", "36047", "36061", "36081", "36085")
#a<-c(unlist(dogs_summary$region))

zip_choropleth(dogs_summary,county_zoom = nyc_fips,
               title       = "Percent spayed or neutered dogs by zip code",
               legend      = "Percentage")
```



The graph above is a spatial heat map of the percent spayed or neutered dogs by zip code. The patterns observed from the graph:
(1)The dogs in “Staten Island” and “Manhattan” boroughs have a higher spayed or neutered percentage;
(2)The dogs in “Queens”, “Brooklyn” and “Bronx” boroughs have a lower spayed or neutered percentage;

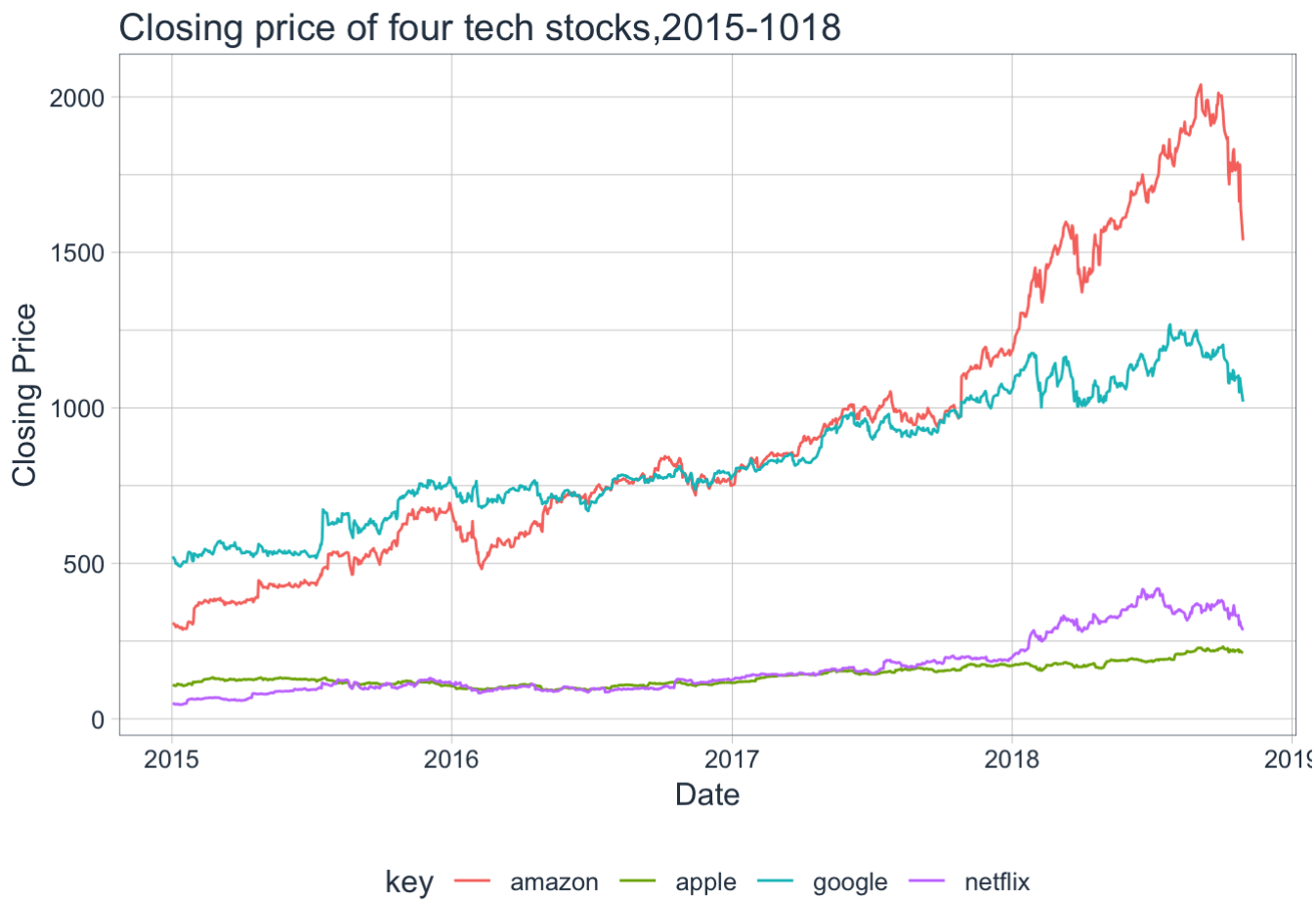
5. Time Series

(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)

# set stock and dates
first.date <- '2015-01-01'
last.date <-  Sys.Date()

# get data with tq_get
apple_stock <- tq_get('AAPL',get = "stock.prices", from = first.date, to = last.date)
amazon_stock <- tq_get('AMZN',get = "stock.prices", from = first.date, to = last.date)
google_stock <- tq_get('GOOG',get = "stock.prices", from = first.date, to = last.date)
netflix_stock <- tq_get('NFLX',get = "stock.prices", from = first.date, to = last.date)
# tidy data
aagn =data.frame(date= apple_stock$date,apple=apple_stock$close , amazon=amazon_stock$cl
ose,google=google_stock$close,netflix=netflix_stock$close)
aagn_tidy<-aagn %>%gather(key,value, -date)
# plot line chart
ggplot(aagn_tidy,aes(x=date, y=value, colour=key)) + geom_line()+labs(title = "Closing p
rice of four tech stocks,2015-1018", y = "Closing Price", x = "Date") + theme_tq(12)
```

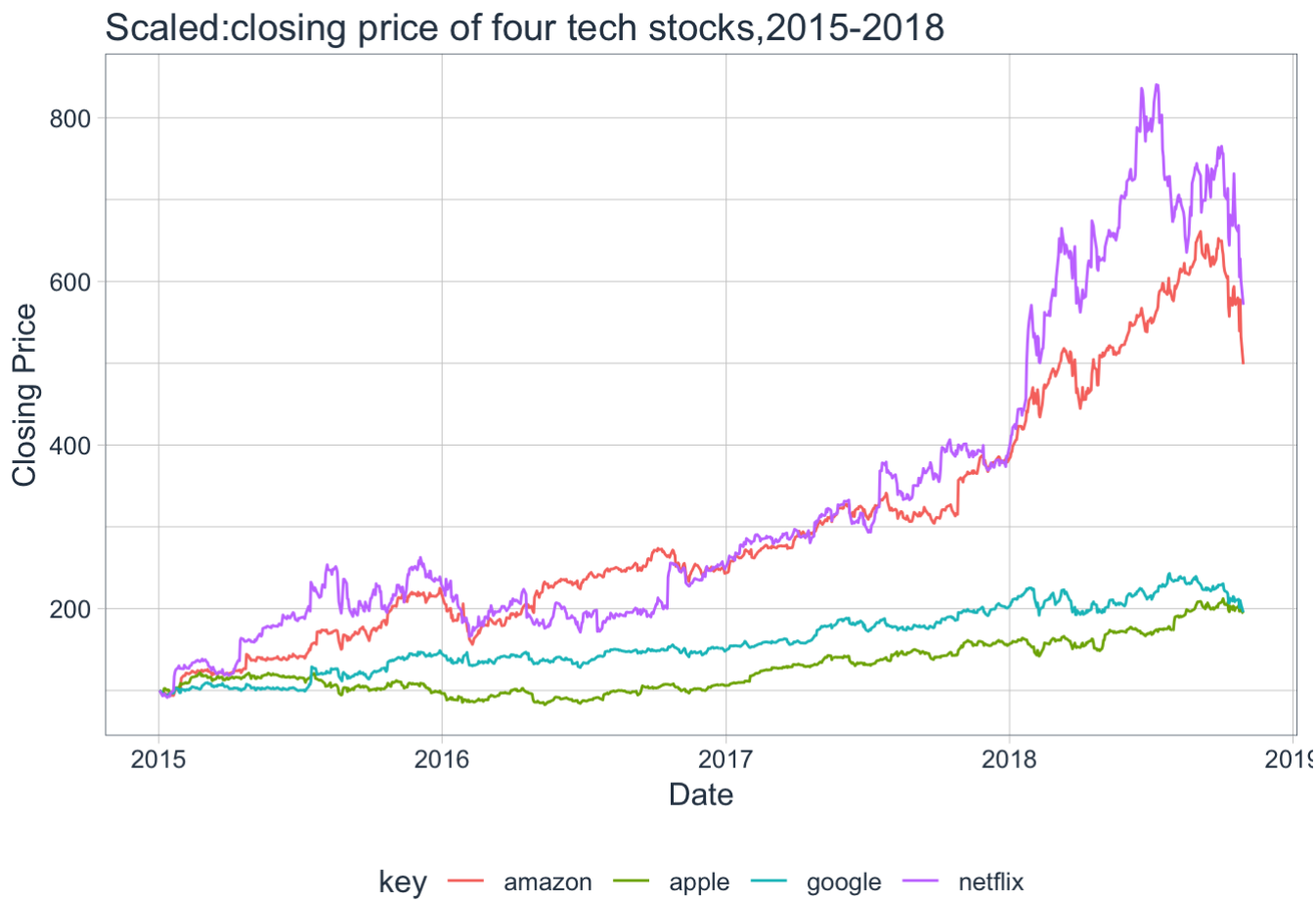


Use the tidyquant package to collect information on four tech stocks - “Apple”, “Google”, “Amazon” and “Netflix” from 2015-01-01 to today. The Line chart of closing price is shown above.

Two groups are shown in the above line plot:
(1)Group 1 -“Amazon” & “Google”: has a higher closing price that over 500 after the year 2016;
(2)Group 2 -“Apple” & “Netflix”: has a lower closing price that below 500 from 2015 up to now.

(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)?

```
aagn_scale <- aagn_tidy %>% group_by(key) %>%
  mutate(index = round(100*value/value[1], 2)) %>%
  ungroup()
# plot line chart
ggplot(aagn_scale,aes(x=date, y=index, colour=key)) + geom_line()+labs(title = "Scaled:c
losing price of four tech stocks,2015-2018", y = "Closing Price", x = "Date") + theme_tq
(12)
```



Scale the data so each stock begins at 100 and replot the line chart shown above. Unlike the line chart in a showing the real stock price with time, this chart plot shows the relative price change compared to the start date.

Two groups are also observed in the scaled line plot:
(1)Group 1 -“Amazon” & “Netflix”: has a higher relative closing price changing with time;
(2)Group 2 -“Apple” & “Google”: has a lower relative closing price changing with time.

From the scaled graph, the changing rate of four tech stocks can be clearly learned, which is not visible in (a).

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. . .)

NYC Health Department, who want to know the time series of total amount born dogs. With the time series of total amount born dogs, it can help the NYC Health Department on environment protection, animal control, and welfares.

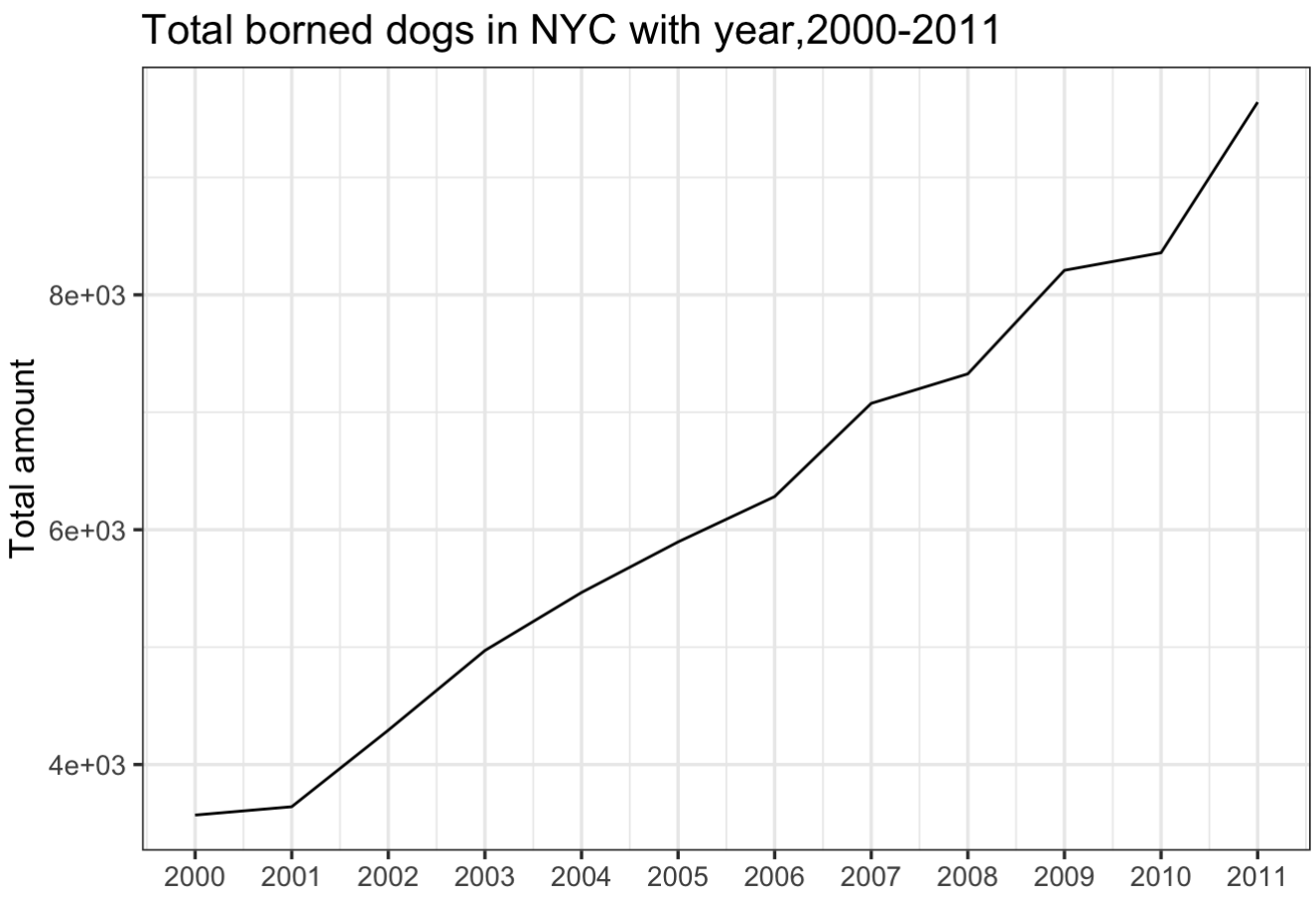
(b) What is the main point you hope someone will take away from the graph?

The graph is a time series plot showing the total born dogs in NYC with the year.

The goal of this time series plot is to find patterns from the NYCdogs data and use these data for predictions, which can help the NYC Health Department on environment protection, animal control, and welfares.

(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.

```
library(scales)
dogs_ts<-dogs%>% group_by(year(dogs$birth)) %>%summarise(n=n())
dogs_ts<-rename(dogs_ts,"year"="year(dogs$birth)")
dogs_ts<-dogs_ts[which(dogs_ts$year>=2000 & dogs_ts$year<=2011),]
# plot line chart
ggplot(dogs_ts,aes(x=as.numeric(year), y=n)) + geom_line()+labs(title = "Total borned do
gs in NYC with year,2000-2011", y ="Total amount", x = "") + scale_x_continuous(breaks=c
(2000:2011),labels=c(2000:2011))+theme_bw(13)+ scale_y_continuous(labels = scientific)
```



The graph above is a time series plot showing the total born dogs in NYC with the year. To have a better representation of the trend of the data, only year 2000-2011 are considered as the data for the year 2012 is incomplete.

From the graph, it can be observed:
(1) the increasing trend with the year of total born dogs in NYC;
(2) the relative stable increasing rate with the year of total born dogs in NYC.

According to the observation above, it's possible that the total born dogs in NYC will keep increasing with the same stable rate with the year. This can help my stakeholder - NYC Health Department for their further environment protection, animal control, and welfare policies.