Homework #3

Xiaowei Zhao

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/

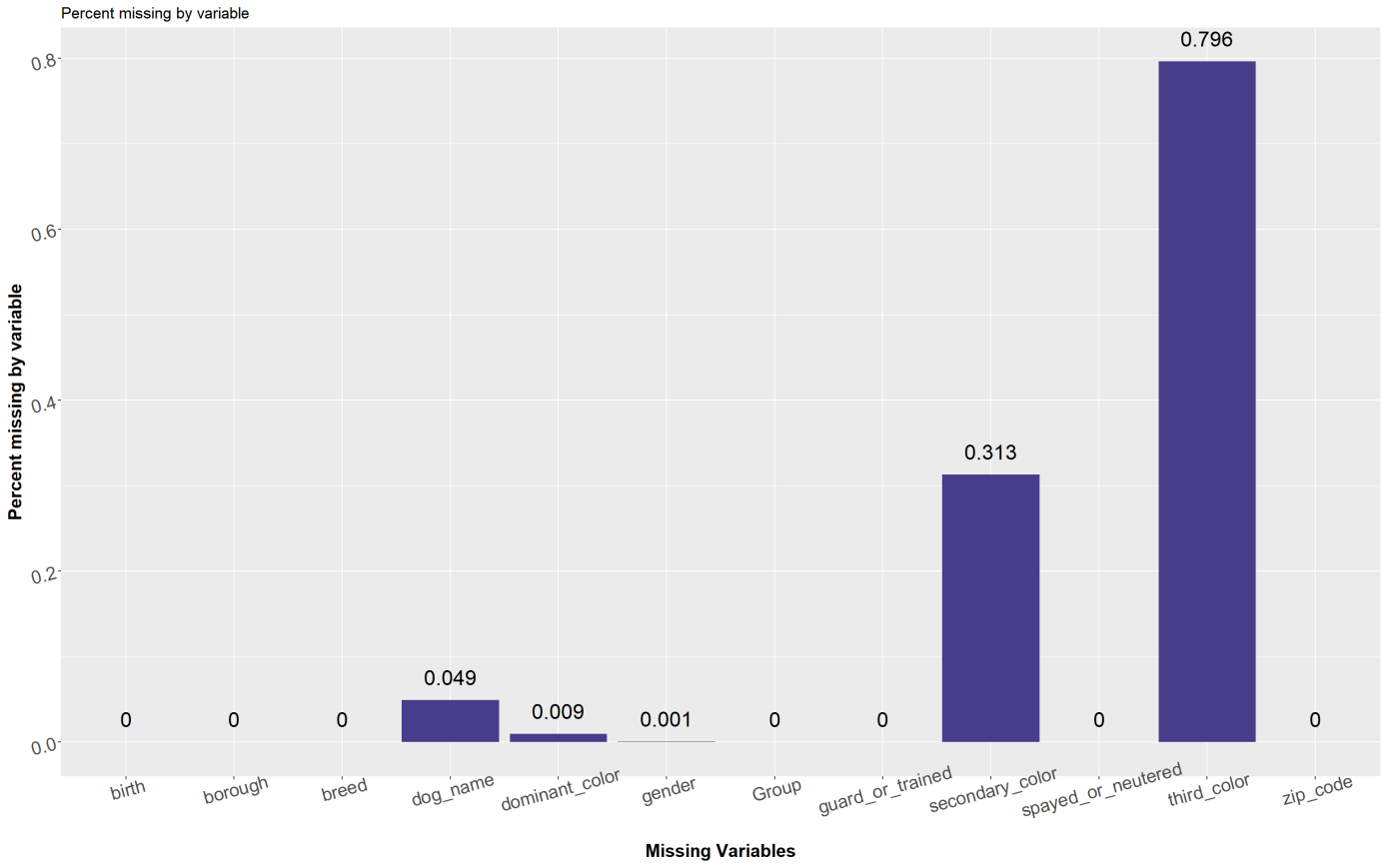
Please use the "NYCdogs.csv" version found in Files/Data folder on CourseWorks, which includes a Group column. If you already did some of the questions that didn't require the Group column, you do not have to redo them.

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). 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/) 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

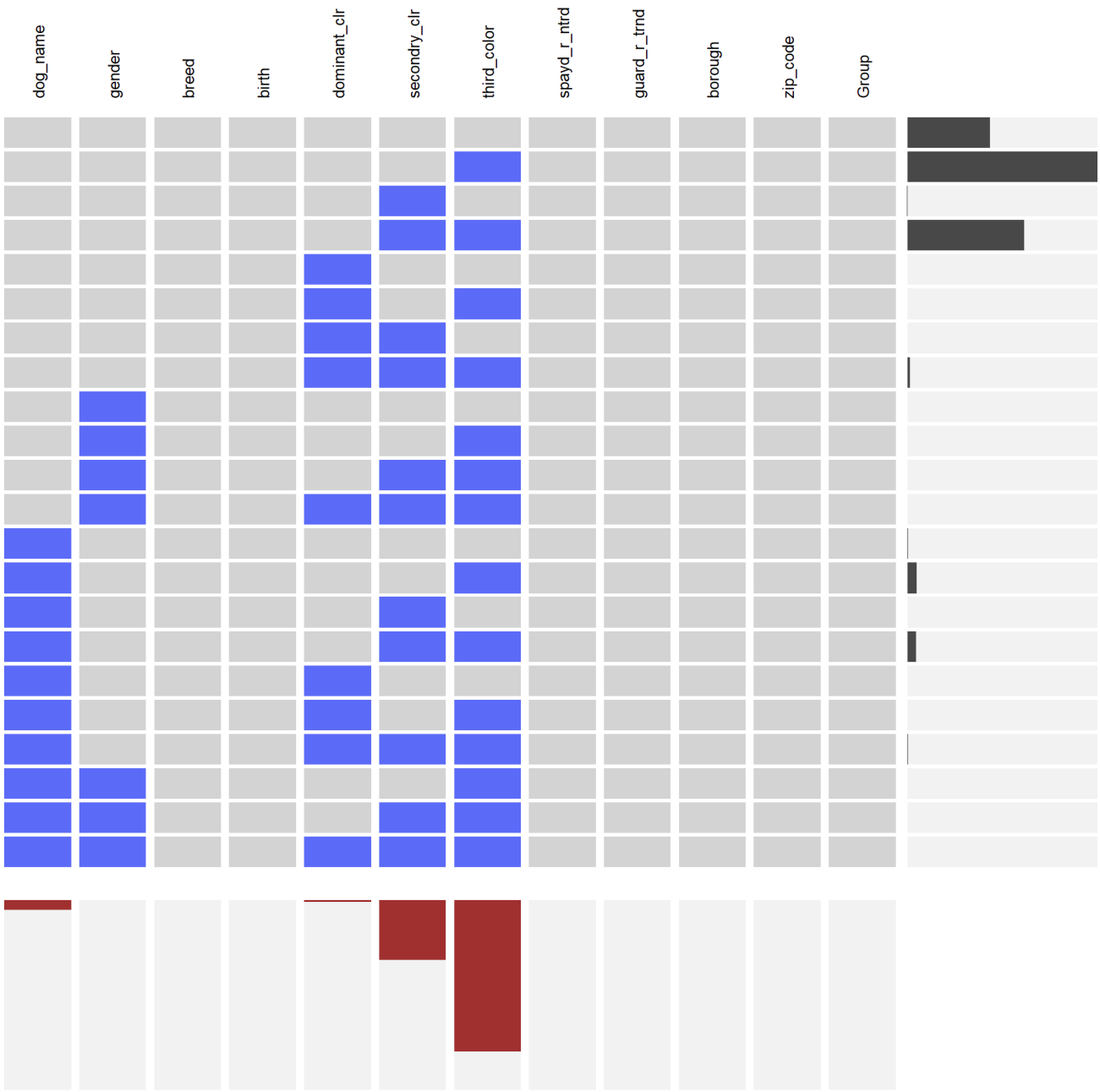
a. Create a bar chart showing percent missing by variable.

```
dogs = read_csv('F:/NYCdogs.csv', na = 'n/a')
Missing_Var=as.data.frame(colSums(is.na(dogs)))
Total_Rows=nrow(dogs)
colnames(Missing_Var)=c('Missing_by_variable')
ggplot(Missing_Var,aes(x=rownames(Missing_Var), y=Missing_by_variable/Total_Rows,label=)) +
  geom_bar(stat  ='identity',fill='darkslateblue') +
  geom_text(aes(label = round(Missing_by_variable/Total_Rows,3),vjust=-1), size = 6) +
  xlab("Missing Variables") +
  ylab("Percent missing by variable") +
  theme(axis.title.x = element_text(size = 15, family = "myFont", face = "bold"),axis.title.y = element_text
(size = 15, family = "myFont", face = "bold"),axis.text=element_text(size=15,vjust = 0.5, hjust = 0.5, angl
e = 15)) +
  ggtitle('Percent missing by variable')
```



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

```
extracat::visna(dogs)
```



There are some variables which have obviously missing patterns. They are third_color, secondry_color and dog_name. Among them, the one which has most missing pattern is third_color. Also, since if a dog doesn't have its third color, it will not have a second color. Hence, we can find a lot of missing values in second_color. As we can see in the graph above, a lot of observations just have the most missing variables in third_color and second most ones in both third_color and second_color.

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

```
percent_missing_gender=dogs %>% group_by(gender) %>%
  summarize(num_dogs = n(), num_na = sum(is.na(dog_name))) %>%
  mutate(percent_na = round(num_na/num_dogs, 2)) %>%
  arrange(-percent_na)
percent_missing_gender
```

```
percent_missing_Group=dogs %>% group_by(Group) %>%
  summarize(num_dogs = n(), num_na = sum(is.na(dog_name))) %>%
  mutate(percent_na = round(num_na/num_dogs, 2)) %>%
  arrange(-percent_na)
percent_missing_Group
```

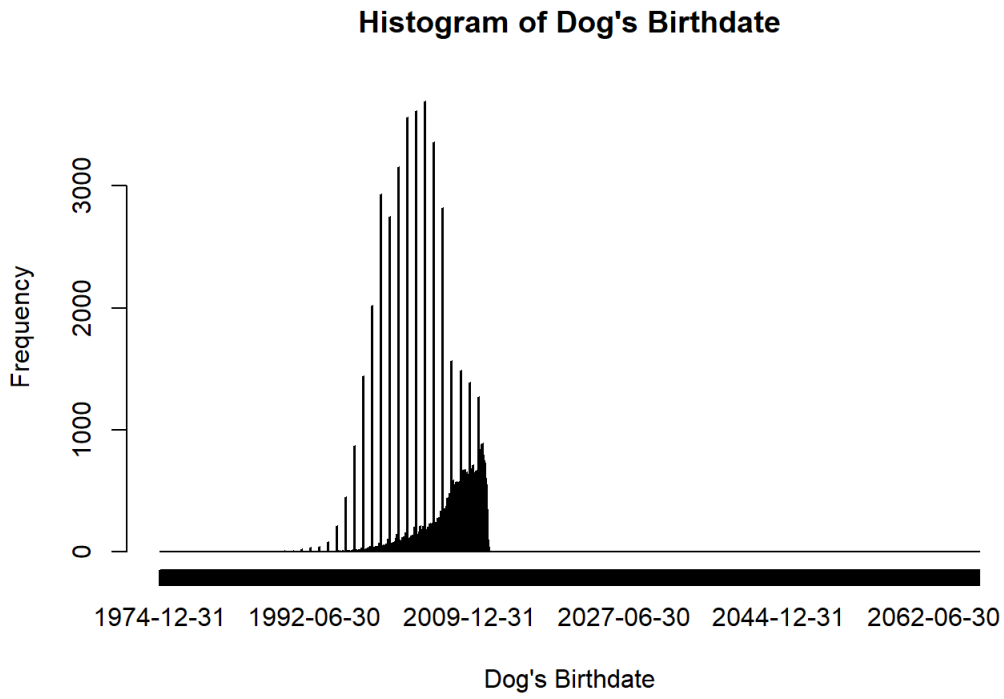
```
percent_missing_borough=dogs %>% group_by(borough) %>%
  summarize(num_dogs = n(), num_na = sum(is.na(dog_name))) %>%
  mutate(percent_na = round(num_na/num_dogs, 2)) %>%
  arrange(-percent_na)
percent_missing_borough
```

As we can see in the above three tables, dog_name missing patterns are not associated with gender and borough because we have similar percentage of NA data here. But it is associated with Group. We can observe a relatively bigger variance among the percentage of NA data. We can find that Herding, Sporting, Terrier dog have low percentage like 0.02 but Non-Sporting dogs have a higer percentage of 0.07.

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.

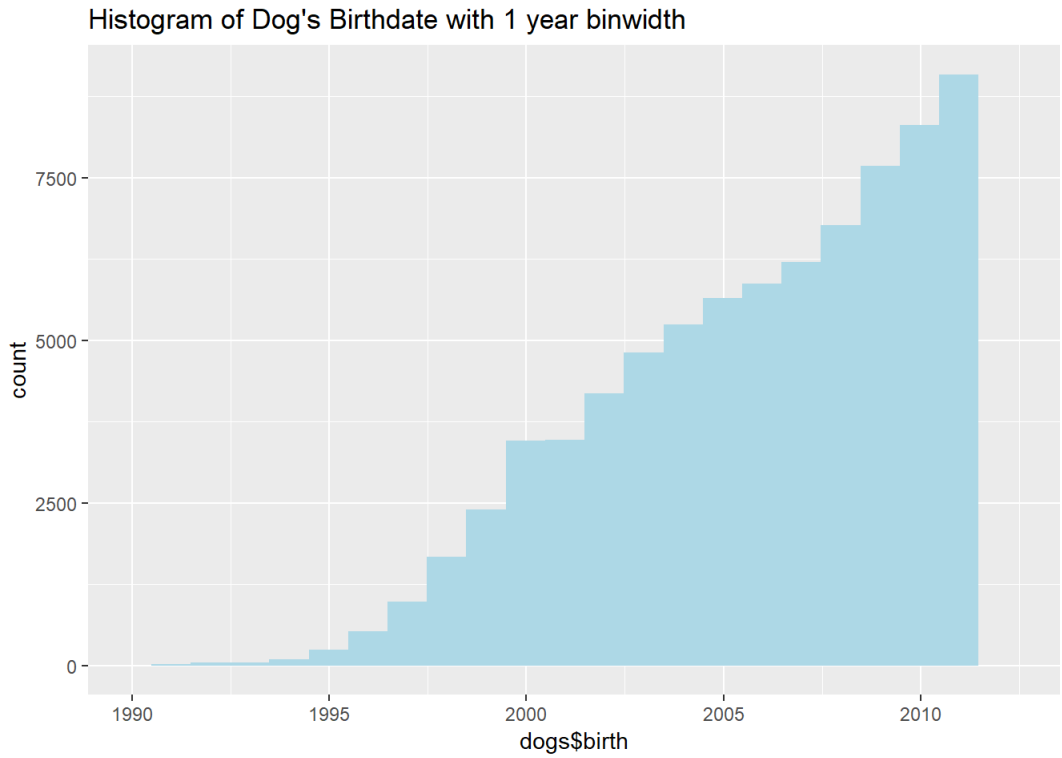
```
dogs$M=str_extract(dogs$birth,'[a-zA-Z]+')
dogs$Y=str_extract(dogs$birth,'\\d+')
dogs$birth=paste(dogs$Y,dogs$M,'01',sep='/')
dogs$birthdate=ymd(dogs$birth)
dogs$birth=as.Date(dogs$birthdate,"%y-%m-%d")
hist(dogs$birth,breaks = "month",freq=TRUE,xlab="Dog's Birthdate")
```



We can make a hypothesis that dogs have very a relatively higher birth rate in some certain months. More specifically, a lot of dogs born in January.

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

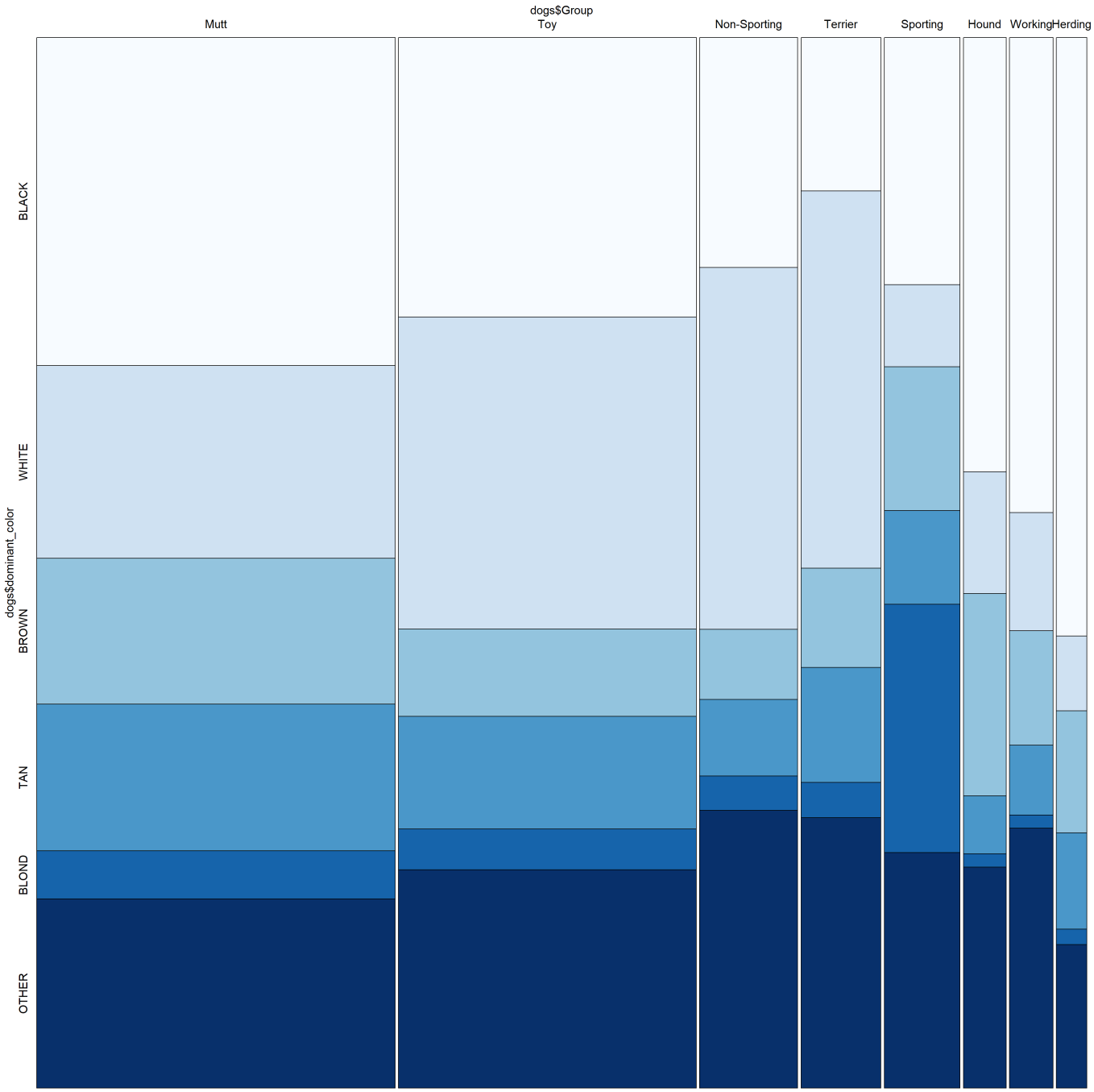
```
ggplot(dogs) +
  geom_histogram(aes(x=dogs$birth), binwidth=365, position="identity",fill='lightblue') +
  scale_x_date(limits = c(ymd("1990-01-01"), ymd("2012-06-01"))) +
  ggtitle("Histogram of Dog's Birthdate with 1 year binwidth")
```



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.

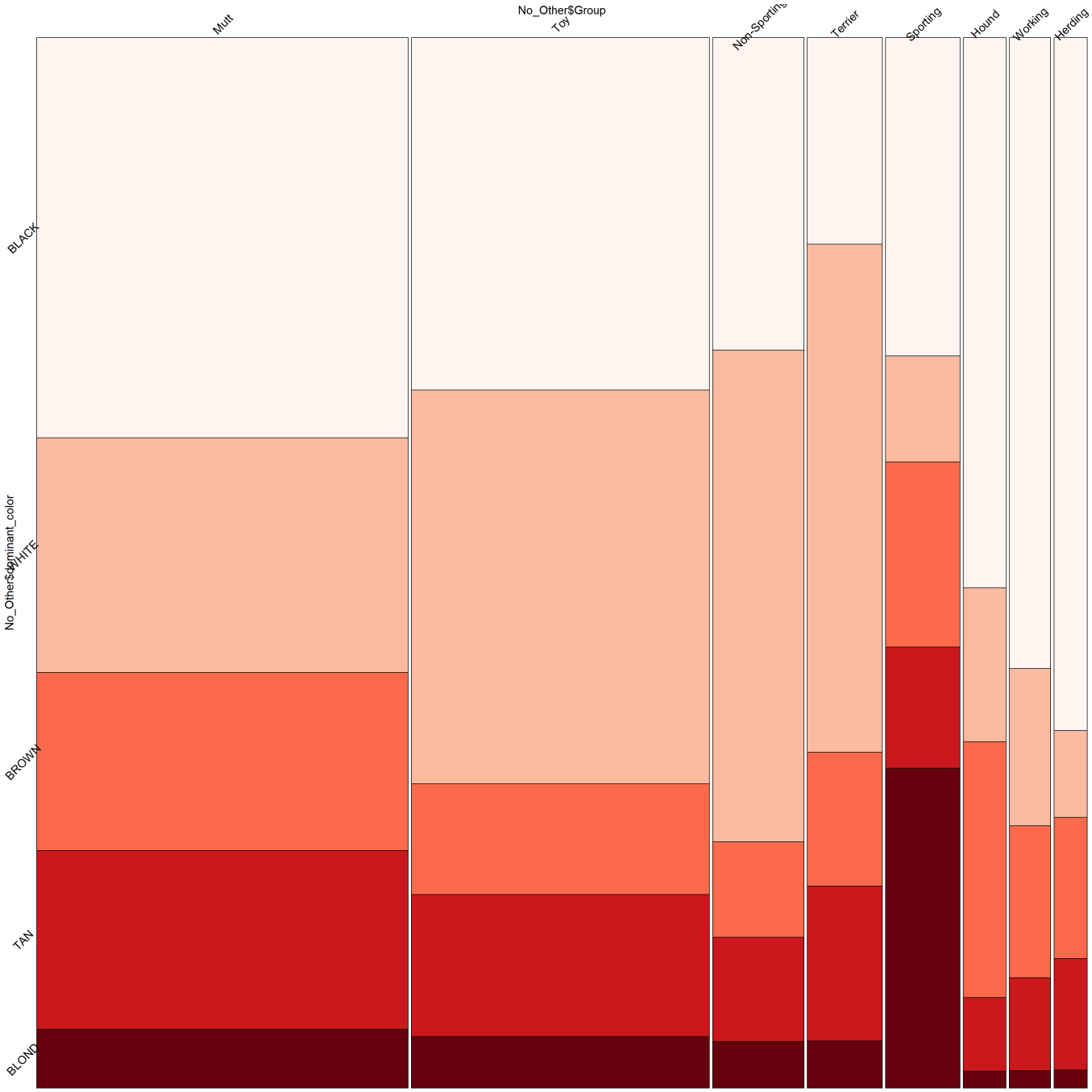
```
temp=dogs %>% group_by(dominant_color) %>% summarise(n_rows=length(dominant_color))
FiveColor=data.frame(temp[order(-temp$n_rows),][1:5,])
dogs$dc=dogs$dominant_color
dogs$dominant_color[!(dogs$dominant_color %in% FiveColor$dominant_color)]="OTHER"
dogs$dominant_color=fct_relevel(dogs$dominant_color,'BLACK','WHITE', 'BROWN','TAN','BLOND','OTHER')
test=dogs %>% group_by(Group) %>% summarise(n_rows=length(Group))
dogs$Group=fct_relevel(dogs$Group,'Mutt','Toy','Non-Sporting','Terrier','Sporting','Hound','Working','Herdin
g')
blues = brewer.pal(9, "Blues")
blue_range = colorRampPalette(blues)
vcd::mosaic(dogs$dominant_color~dogs$Group,direction=c('v','h'),gp=gpar(fill=blue_range(6)),labeling_args=li
st(gp_labels=gpar(fontsize=15),gp_varnames=gpar(fontsize=15)))
```



There are some patterns to show that variables appear to be associated. For example, as we can see in the graph, black is the most dominant color, if there are no associations between variables, we should see a pattern that black is always the dominant color among all dog groups. Like both Non-sporting and Terrier dogs have dominant color in white not black. Also, blond is the least dominant color among top five color, however, we can observe from the graph that Sporting dog have dominant color in blond which is not following the order of dominant color.

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?

```
No_Other=dogs
No_Other=filter(No_Other,No_Other$dominant_color!='OTHER')
reds = brewer.pal(9, "Reds")
red_range = colorRampPalette(reds)
No_Other$dominant_color=fct_relevel(No_Other$dominant_color,'BLACK','WHITE', 'BROWN','TAN','BLOND')
No_Other$dominant_color=droplevels(No_Other$dominant_color)
No_Other$Group=fct_relevel(No_Other$Group,'Mutt','Toy','Non-Sporting','Terrier','Sporting','Hound','Working'
,'Herding')
vcd::mosaic(No_Other$dominant_color~No_Other$Group,direction=c('v','h'),rot_labels=c(45,90,45,45),gp=gpar(fi
ll=red_range(5)),labeling_args=list(gp_labels=gpar(fontsize=15),gp_varnames=gpar(fontsize=15)))
```

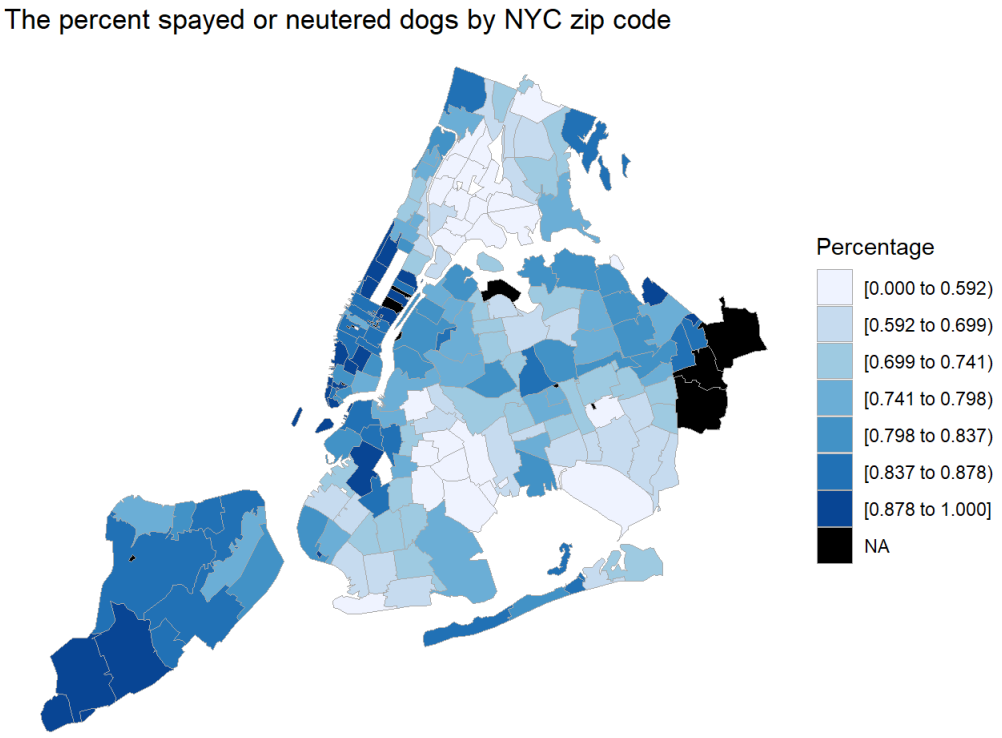


The result don't change except that we just leave the OTHER. However, the graph above Without OTHER is more clear to observe the patterns than previous one. Because here we have five dominant colors more proportion to the color, we will have a better graph to find patterns. But if in another situation that we exclude OTHER, we oberserve that the associations and patterns among variables has changed(like relative frequency), we should include OTHER. Also, here we don't need information in OTHER so we can exclude it. If in another situation, we need to know some information and patterns in OTHER, we cannot exclude it.

4. Maps

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

```
SN=table(dogs$zip_code,dogs$spayed_or_neutered)
SN_df=as.data.frame.matrix(SN)
percent=SN_df$Yes/(SN_df$No+SN_df$Yes)
SN_df$percent=percent
SN_df2=data.frame(region =rownames(SN_df),value=SN_df$percent)
nyc_fips = c(36005, 36047, 36061, 36081, 36085)
zip_choropleth(SN_df2,county_zoom = nyc_fips,title='The percent spayed or neutered dogs by NYC zip code',leg
end='Percentage')
```

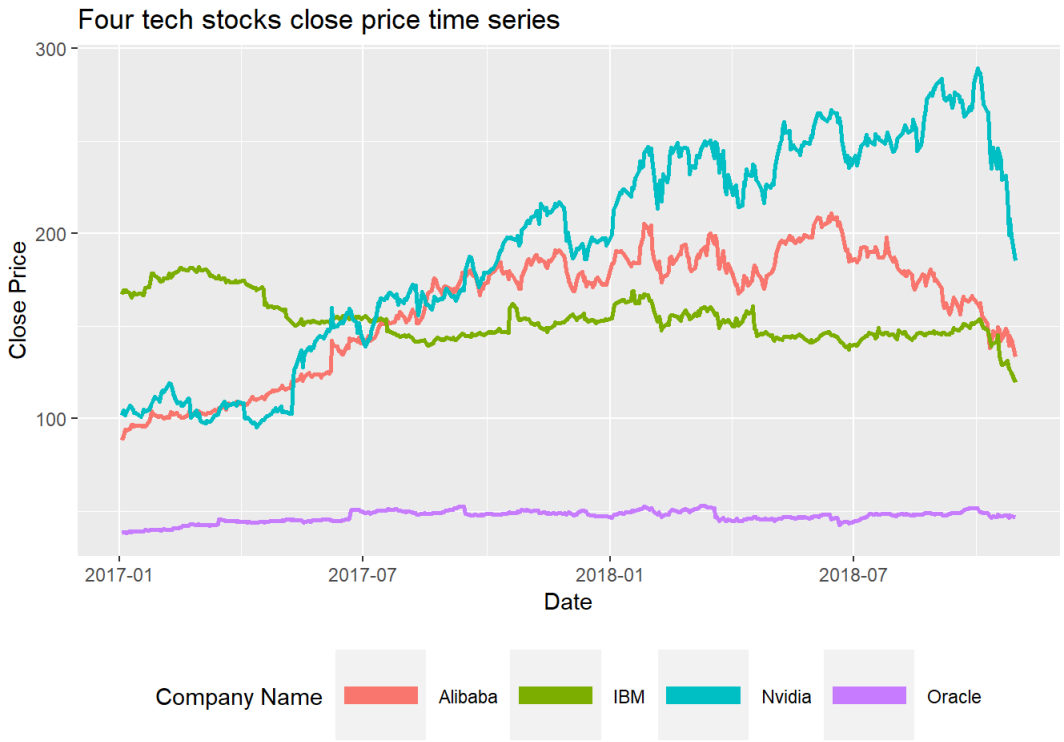


The percentage of spayed or neutered dogs are higher in lower Manhattan, midtown and State Island. It is relatively lower in upper Manhattan, Bronx and Brooklyn

5. Time Series

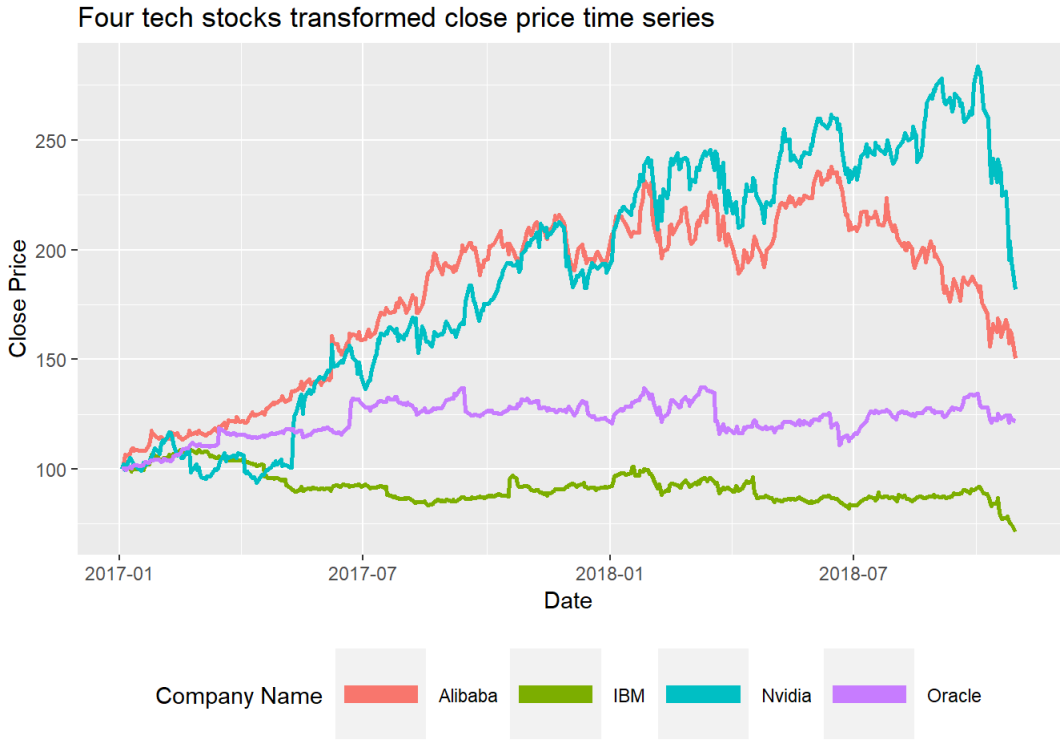
a. Use the `tidyquant` package to collect information on four tech stocks of your choosing. Create a multiple line chart of the closing prices of the four stocks on the same graph, showing each stock in a different color.

```
price = c('NVDA', 'IBM','BABA','ORCL') %>% tq_get(get  = "stock.prices",from='2017-01-01')
price$symbol=recode(price$symbol,'NVDA'='Nvidia','IBM'='IBM','BABA'='Alibaba','ORCL'='Oracle')
ggplot(price,aes(x=date,y=close,color=symbol)) +
  geom_line(size=1) +
  xlab('Date') +
  ylab('Close Price') +
  theme(legend.position="bottom") +
  theme(legend.key.size = unit(3,"line")) +
  guides(colour = guide_legend(override.aes = list(size=4))) +
  scale_color_discrete(name = "Company Name") +
  ggtitle('Four tech stocks close price time series')
```



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)?

```
price1 = price %>% group_by(symbol) %>% mutate(start=first(close),trans_price=(close/start)*100)
ggplot(price1,aes(x=date,y=trans_price,color=symbol)) +
  geom_line(size=1) +
  xlab('Date') +
  ylab('Close Price') +
  theme(legend.position="bottom") +
  theme(legend.key.size = unit(3,"line")) +
  guides(colour = guide_legend(override.aes = list(size=4))) +
  scale_color_discrete(name = "Company Name") +
  ggtitle('Four tech stocks transformed close price time series')
```



After transforming the close price to the same start price, we can oberserve the correlation and trend of these four stocks more clearly. Oracle have relatively the lowest price among all 4 stocks but after the transformation, its performance is better than IBM which has a much more higher start price than Oracle. We can also find that Nvidia has the best performance among all 4 stocks

although its start price is not the highest. Moreover, Nvidia-Alibaba and IBM-Oracle have similar trend.

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…)

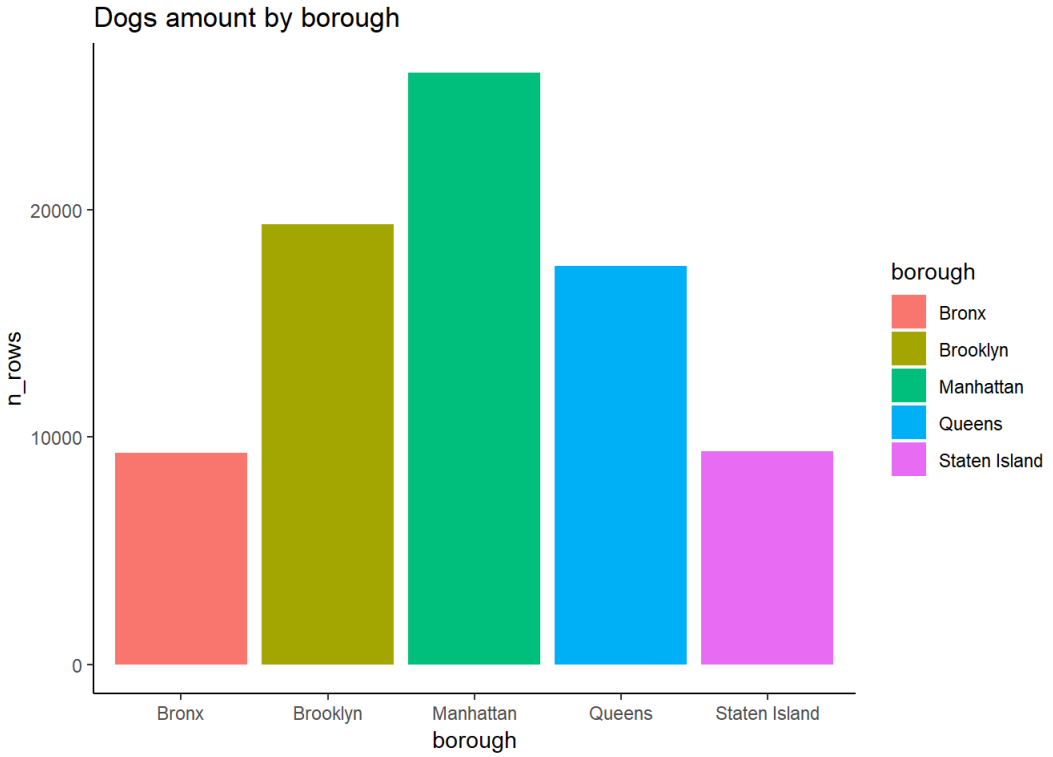
A pets related start-up company wants to set up several pet stores in NYC. They want to know where to locate their stores in order to make more profits

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

The main point is that to find which area or neighborhood has realtively higher dogs amount so that the company can be condient in setting their pets stores in those selected region to make potentially bigger profits.

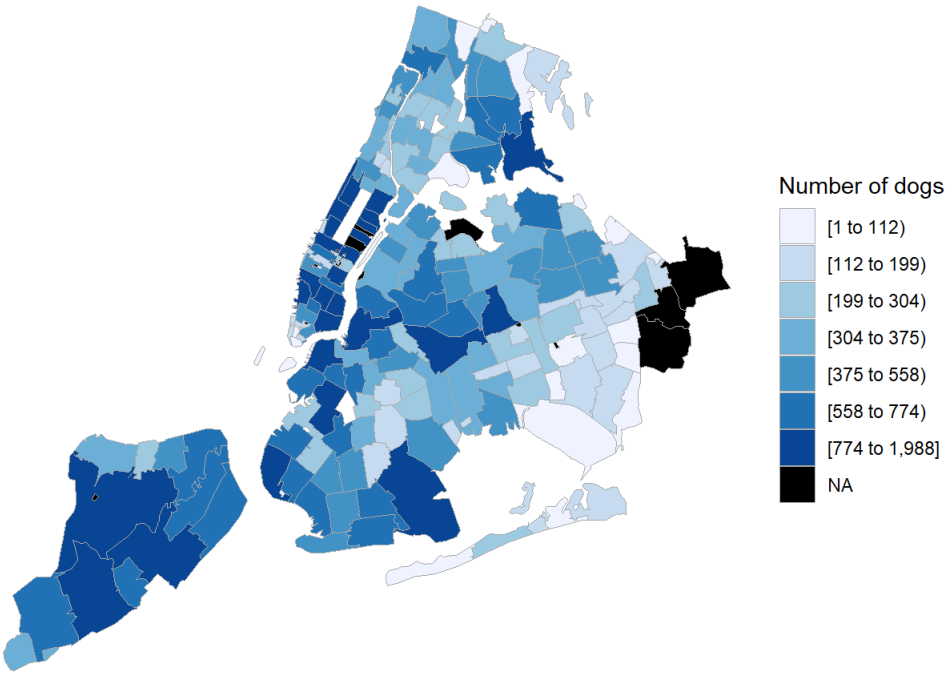
c. Present the graph, cleaned up to the standards of “presentation style.” Pay attention to choice of graph type, if and how the data will be summarized, if and how the data will be subsetted, title, axis labels, axis breaks, axis tick mark labels, color, gridlines, and any other relevant features.

```
mydata=dogs %>% group_by(borough) %>% summarise(n_rows=length(borough))
ggplot(mydata,aes(x=borough,y=n_rows,fill=borough)) +
  geom_bar(stat='identity') +
  ggtitle('Dogs amount by borough') +
  theme_classic()
```




```
n_dogs=dogs %>% group_by(zip_code) %>% summarize(num_dogs = n())
mydf=data.frame(region = as.character(n_dogs$zip_code),value=n_dogs$num_dogs)
nyc_fips = c(36005, 36047, 36061, 36081, 36085)
zip_choropleth(mydf,county_zoom = nyc_fips,title='The amount of dogs by NYC zip code',legend='Number of dogs
')
```

The amount of dogs by NYC zip code



Because of the prossible limitation of budget and consideration of any kind of cost, the company may consider setting up their precious pets stores in Manhhattan from the result of the first graph, since Manhattan has realtively highest dogs amount. Also, they will focus on lower Manhattan and midtown to spread their business because those areas have relatively higher dogs amount than other regions in Manhattan.