Dog_vs_Cat

Bowen Xiao

May 27, 2018

Dataset introduction

The city of Seattle makes available its database of pet licenses issued from 2005 to the beginning of 2017 as part of the city's ongoing Open Data Initiative. The data is also obtainable from the Socrata Open Data Access (SODA) portal in either CSV or JSON formats.

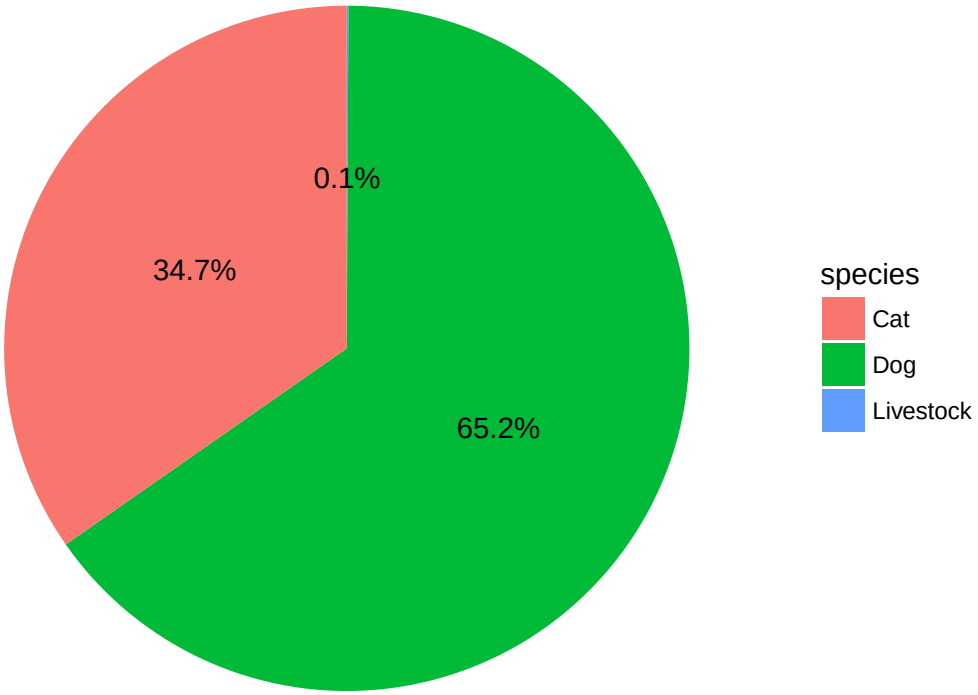
```
library(readr)
licenses <- read_csv("seattle_pet_licenses.csv")
dog<-licenses[licenses$species=='Dog',]
cat<-licenses[licenses$species=='Cat',]
library(knitr)
kable(data.frame(attributes=colnames(licenses)))
```

attributes
animal_s_name
license_issue_date
license_number
primary_breed
secondary_breed
species
zip_code

Most common species

Not surprisingly, most registered pets in Seattle are either dog or cat, and there are more dogs than cats. Thus I am going to focus on dogs and cats in the following.

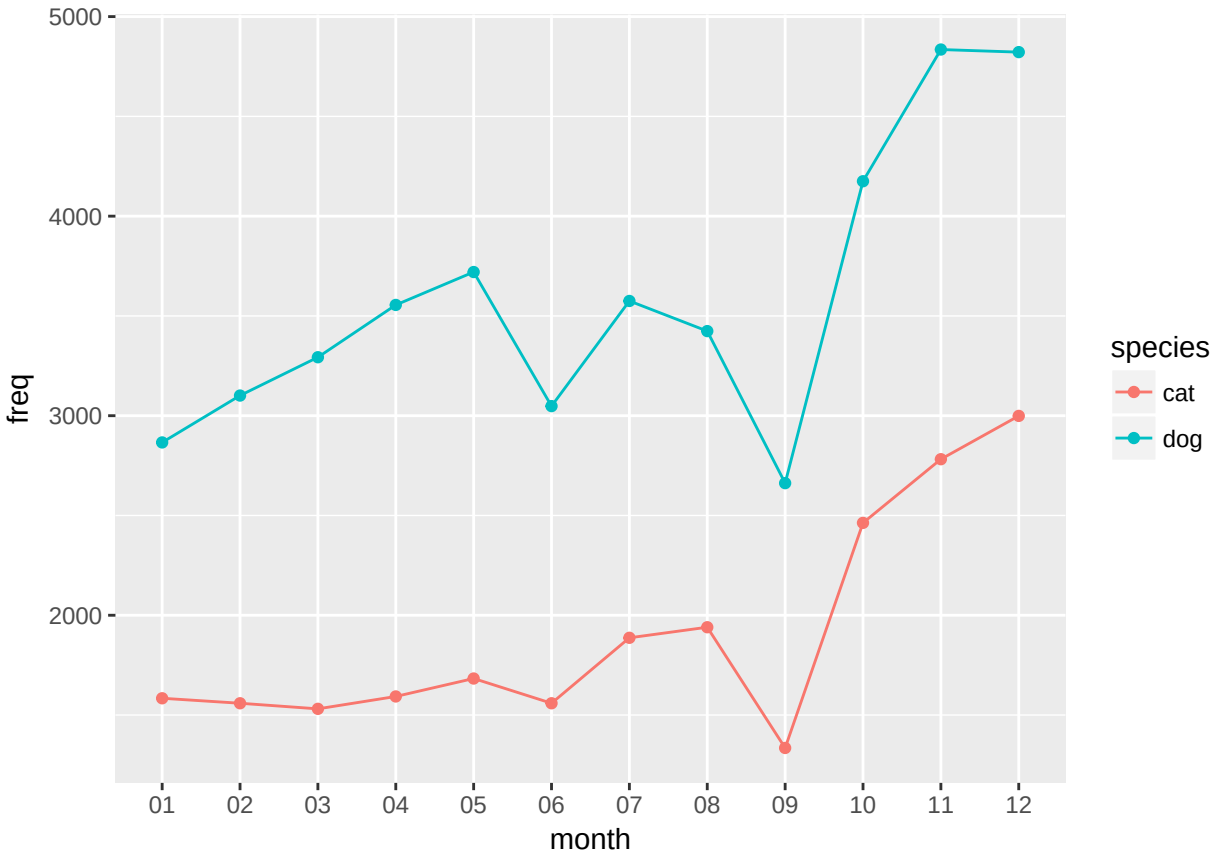
```
library(ggplot2)
library(dplyr)

data <- licenses %>%
  group_by(species) %>%
  count() %>%
  ungroup() %>%
  mutate(per=`n`/sum(`n`)) %>%
  arrange(desc(species))
data$label <- scales::percent(data$per)

ggplot(data=data)+
  geom_bar(aes(x="", y=per, fill=species), stat="identity", width = 1)+
  coord_polar("y", start=0)+
  theme_void()+
  geom_text(aes(x=1, y = cumsum(per) - per/2, label=label))
```



Most popular month for registeration

Most people chosed to registered their pets in end of a year.

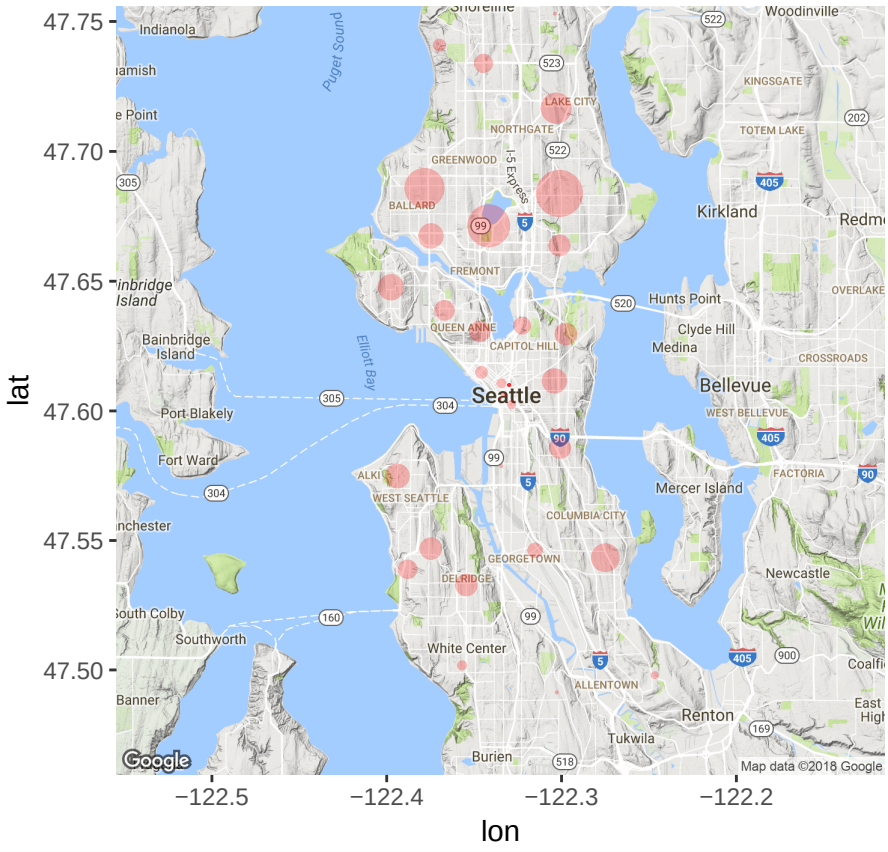
```
dog.date<-data.frame(date=(format(as.Date(dog$license_issue_date),"%m")))
dog.date<-dog.date[complete.cases(dog.date),]
dog.date<-as.factor(dog.date)
dog.date<-as.data.frame(table(dog.date))
colnames(dog.date)=c('month','freq')
dog.date$species<-rep('dog',nrow(dog.date))

cat.date<-data.frame(date=(format(as.Date(cat$license_issue_date),"%m")))
cat.date<-cat.date[complete.cases(cat.date),]
cat.date<-as.factor(cat.date)
cat.date<-as.data.frame(table(cat.date))
colnames(cat.date)=c('month','freq')
cat.date$species<-rep('cat',nrow(cat.date))

date<-rbind(dog.date,cat.date)
date<-date[order(date$month),]

(p1 <- ggplot(data = date, aes(x = month, y = freq, colour = species)) +
    geom_line(aes(group = species)) +
    geom_point())
```



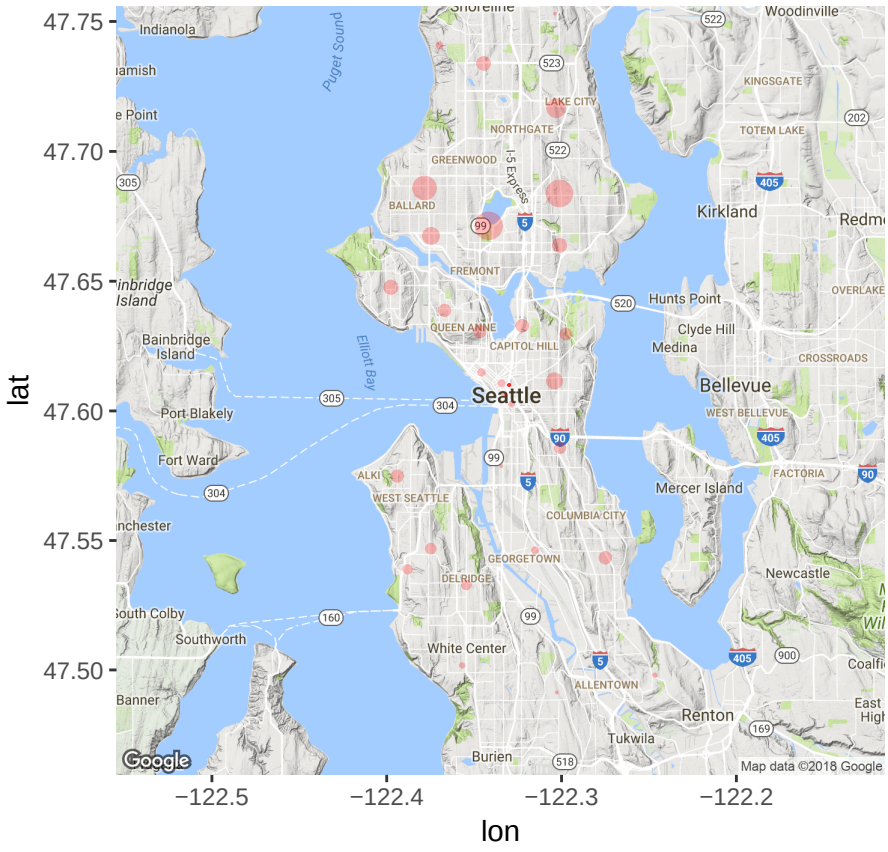
Addresses of pets

Both dogs and cats are everywhere, and it is true that dogs are always outnumber cats in Seattle.

```
library(zipcode)
library(ggmap)

data(zipcode)
seattle<-zipcode[zipcode$city=='Seattle',]

p <- ggmap(get_googlemap(center = c(lon = -122.335167, lat = 47.608013),
                                                    zoom = 11, scale = 2,
                                                    maptype ='terrain',
                                                    color = 'color'))


dog.zip<-dog$zip_code
dog.zip<-dog.zip[complete.cases(dog.zip)]
dog.zip<-as.factor(dog.zip)
dog.zip<-as.data.frame(table(dog.zip))
colnames(dog.zip)=c('zip','freq')
dog.zip<-merge(x = seattle, y = dog.zip, by = "zip", all.x = TRUE)
p + geom_point(aes(x = longitude, y = latitude), data = dog.zip, col='red',size = dog.zip$freq/500,alpha
  theme(legend.position="bottom")
```




```
cat.zip<-cat$zip_code
cat.zip<-cat.zip[complete.cases(cat.zip)]
cat.zip<-as.factor(cat.zip)
cat.zip<-as.data.frame(table(cat.zip))
colnames(cat.zip)=c('zip','freq')
cat.zip<-merge(x = seattle, y = cat.zip, by = "zip", all.x = TRUE)
p + geom_point(aes(x = longitude, y = latitude), data = cat.zip, col='red',size = cat.zip$freq/500,alpha
  theme(legend.position="bottom")
```

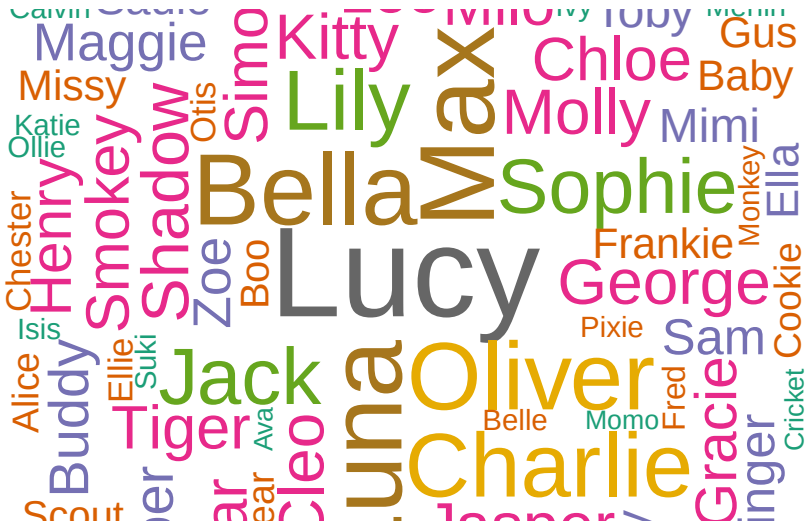


Most popular names for pets

'Lucy', 'Charlie', 'Bella', 'Luna', 'Oliver' and 'Max' seem to be shared names for dogs and cats. However, 'Buddy' and 'Sadie' seem to only works for dogs.

```
library(wordcloud)

dog.names<-dog$animal_s_name
dog.names<-dog.names[complete.cases(dog.names)]
dog.names<-as.data.frame(table(dog.names))
wordcloud(words = dog.names$dog.names, freq = dog.names$Freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```


```
cat.names<-cat$animal_s_name
cat.names<-cat.names[complete.cases(cat.names)]
cat.names<-as.data.frame(table(cat.names))
wordcloud(words = cat.names$cat.names, freq = cat.names$Freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



Conclusion

As is shown above, there is a roughly '4-6' principle for dogs and cats in Seattle when it comes to registration time, addresses and names. It is really hard to guess statistically whether a registered pet is a dog or cat in Seattle.