Lab 2

[redacted]

2023-02-06

```
### This lab will prepare you for PS4!
### This is a _graded lab_, you'll get 3
### points if you solve and submit it correctly

### Squrrel census data is downloaded from NY Open Data Portal
### See the readme
### https://bitbucket.org/otoomet/data/src/master/nature/
### for more information and better layout of the
### variable names
###
### Variables:
###
### **X**: Longitude coordinate for squirrel sighting point
### **Y**: Latitude coordinate for squirrel sighting point
### **Unique Squirrel ID**: Identification tag for each squirrel
### sightings. The tag is comprised of "Hectare ID" + "Shift" + "Date" +
### "Hectare Squirrel Number."
### **Hectare**: ID tag, which is derived from the hectare grid used to
### divide and count the park area. One axis that runs predominantly
### north-to-south is numerical (1-42), and the axis that runs
### predominantly east-to-west is roman characters (A-I).
### **Shift**: Value is either "AM" or "PM," to communicate whether or
### not the sighting session occurred in the morning or late afternoon.
### **Date**: Concatenation of the sighting session day and month.
### **Hectare Squirrel Number**: Number within the chronological
### sequence of squirrel sightings for a discrete sighting session.
### **Age**: Value is either "Adult" or "Juvenile."
### **Primary Fur Color**: Value is either "Gray," "Cinnamon" or
### "Black."
### **Highlight Fur Color**: Discrete value or string values comprised
### of "Gray," "Cinnamon" or "Black."
### **Combination of Primary and Highlight Color**: A combination of the
### previous two columns; this column gives the total permutations of
### primary and highlight colors observed.
### **Color notes**: Sighters occasionally added commentary on the
### squirrel fur conditions. These notes are provided here.
### **Location**: Value is either "Ground Plane" or "Above Ground."
### Sighters were instructed to indicate the location of where the
### squirrel was when first sighted.
### **Above Ground Sighter Measurement**: For squirrel sightings on the
### ground plane, fields were populated with a value of "FALSE."
### **Specific Location**: Sighters occasionally added commentary on the
### squirrel location. These notes are provided here.
### **Running**: Squirrel was seen running.
### **Chasing**: Squirrel was seen chasing another squirrel.
### **Climbing**: Squirrel was seen climbing a tree or other
### environmental landmark.
### **Eating**: Squirrel was seen eating.
### **Foraging**: Squirrel was seen foraging for food.
### **Other Activities**:
### **Kuks**: Squirrel was heard kukking, a chirpy vocal communication
### used for a variety of reasons.
### **Quaas**: Squirrel was heard quaaing, an elongated vocal
### communication which can indicate the presence of a ground predator
### such as a dog.
### **Moans**: Squirrel was heard moaning, a high-pitched vocal
### communication which can indicate the presence of an air predator
### such as a hawk.
### **Tail flags**: Squirrel was seen flagging its tail. Flagging is a
### whipping motion used to exaggerate squirrel's size and confuse
### rivals or predators. Looks as if the squirrel is scribbling with
### tail into the air.
### **Tail twitches**: Squirrel was seen twitching its tail. Looks like
### a wave running through the tail, like a breakdancer doing the arm
### wave. Often used to communicate interest, curiosity.
### **Approaches**: Squirrel was seen approaching human, seeking food.
### **Indifferent**: Squirrel was indifferent to human presence.
### **Runs from**: Squirrel was seen running from humans, seeing them as
### a threat.
### **Other Interactions**: Sighter notes on other types of interactions
### between squirrels and humans.
### **Lat/Long**: Latitude and longitude

## Load tidyverse (or dplyr) library
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Load the dataset
df <- read.csv('nyc-central-park-squirrel-census-2019.csv.bz2')

## How many rows and columns does it contain?
nrow(df)
```

```
## [1] 3023
```

```
ncol(df)
```

```
## [1] 31
```

```
## What are the variable names?
colnames(df)
```

```
##  [1] "X"
##  [2] "Y"
##  [3] "Unique.Squirrel.ID"
##  [4] "Hectare"
##  [5] "Shift"
##  [6] "Date"
##  [7] "Hectare.Squirrel.Number"
##  [8] "Age"
##  [9] "Primary.Fur.Color"
## [10] "Highlight.Fur.Color"
## [11] "Combination.of.Primary.and.Highlight.Color"
## [12] "Color.notes"
## [13] "Location"
## [14] "Above.Ground.Sighter.Measurement"
## [15] "Specific.Location"
## [16] "Running"
## [17] "Chasing"
## [18] "Climbing"
## [19] "Eating"
## [20] "Foraging"
## [21] "Other.Activities"
## [22] "Kuks"
## [23] "Quaas"
## [24] "Moans"
## [25] "Tail.flags"
## [26] "Tail.twitches"
## [27] "Approaches"
## [28] "Indifferent"
## [29] "Runs.from"
## [30] "Other.Interactions"
## [31] "Lat.Long"
```

```
## Show a few lines of data!
head(df, 3)
```

```
##           X        Y Unique.Squirrel.ID Hectare Shift     Date
## 1 -73.95613 40.79408     37F-PM-1014-03     37F    PM 10142018
## 2 -73.96886 40.78378     21B-AM-1019-04     21B    AM 10192018
## 3 -73.97428 40.77553     11B-PM-1014-08     11B    PM 10142018
##   Hectare.Squirrel.Number Age Primary.Fur.Color Highlight.Fur.Color
## 1                       3
## 2                       4
## 3                       8                  Gray
##   Combination.of.Primary.and.Highlight.Color Color.notes     Location
## 1                                          +
## 2                                          +
## 3                                      Gray+             Above Ground
##   Above.Ground.Sighter.Measurement Specific.Location Running Chasing Climbing
## 1                                                      false   false    false
## 2                                                      false   false    false
## 3                               10                     false    true    false
##   Eating Foraging Other.Activities  Kuks Quaas Moans Tail.flags Tail.twitches
## 1  false    false                  false false false      false         false
## 2  false    false                  false false false      false         false
## 3  false    false                  false false false      false         false
##   Approaches Indifferent Runs.from Other.Interactions
## 1      false       false     false
## 2      false       false     false
## 3      false       false     false
##                                      Lat.Long
## 1 POINT (-73.9561344937861 40.7940823884086)
## 2 POINT (-73.9688574691102 40.7837825208444)
## 3 POINT (-73.97428114848522 40.775533619083)
```

```r
## How many different unique squirrels are there?
n_distinct(df$Unique.Squirrel.ID)
```

```
## [1] 3018
```

```r
## How many squirrels were Approaching humans?
num.approach <- nrow(df[df$Approaches=='true',])
num.approach
```

```
## [1] 178
```

```r
## How many squirrels were indifferent, and how many
## were running from humans?
## use a single 'summarize()' to compute it!
df %>% summarize(
    indifferent = sum(ifelse(Indifferent=='true', 1, 0)),
    runs.from = sum(ifelse(Runs.from=='true', 1, 0))
  )
```

```
##   indifferent runs.from
## 1        1454       678
```

```r
## Compute percentage of squirrels who are approaching
## humans
perc.approach <- num.approach / nrow(df)
perc.approach
```

```
## [1] 0.05888191
```

```r
## Show 10 randomly selected 'Other Activities' what squirrels do
## But only if those are not NA
df %>% filter(Other.Activities != '') %>% sample_n(10) %>% select(Other.Activities)
```

```
##                                                 Other.Activities
## 1                                                       sitting
## 2                             stole (found?) an entire sandwich
## 3                    made a back-door escape from dog off-leash
## 4                                                      watching
## 5                                        carrying food in mouth
## 6                                                       digging
## 7                                                       playing
## 8                                         chasing (#7),playing?
## 9                                    burying the food on ground
## 10 stood still & watched me then jumped on a fence and ran away
```

```
## Are there any squirrels who are climbing to
## approach humans?
df %>% filter(Climbing=='true' & Approaches=='true') %>% count()
```

```
##    n
## 1 27
```

```
# Yes, there are 27 squirrels who are climbing to approach humans.

## What kind of values are there for squirrel age?
unique(df$Age)
```

```
## [1] ""         "Adult"    "Juvenile" "?"
```

```
## Explain what you see.  What does it tell about data quality?
# For squirrel age, We have missings, coded in two ways (NA and '?').
# Otherwise looks good but not particularly precise with only two age categories (Adult, Juvenile).
```

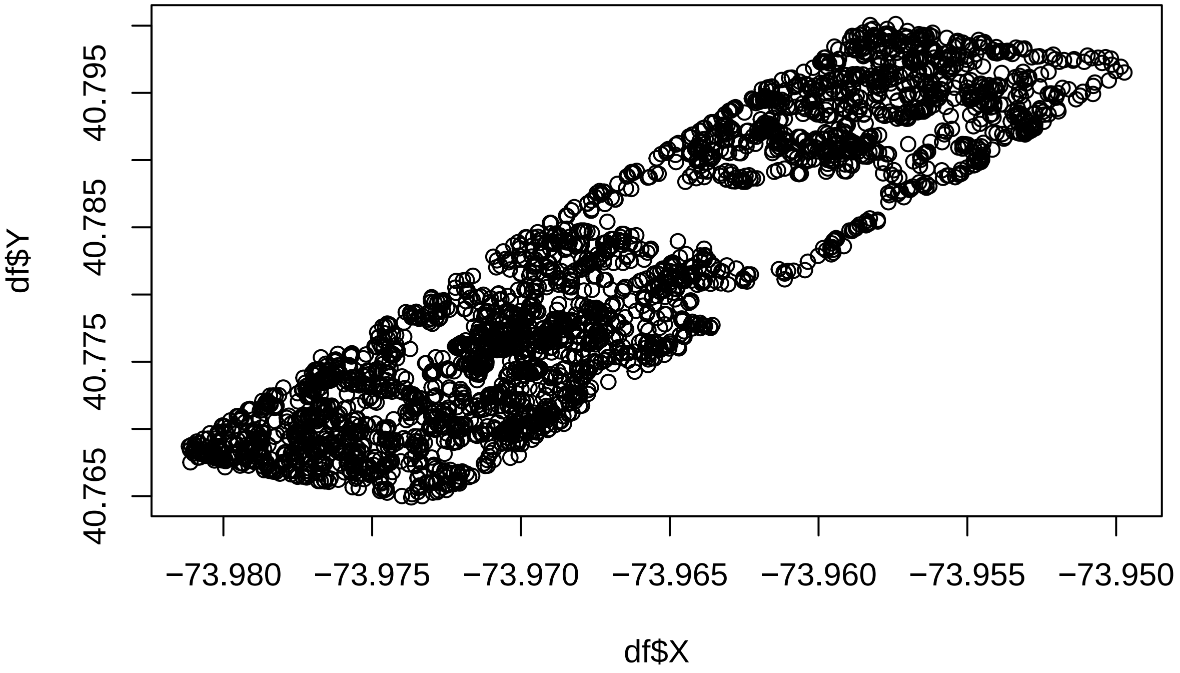
```
## How many squirrels of different age group were observed?
df %>% group_by(Age) %>% count()
```

```
## # A tibble: 4 x 2
## # Groups:   Age [4]
##   Age            n
##   <chr>      <int>
## 1 ""           121
## 2 "?"            4
## 3 "Adult"     2568
## 4 "Juvenile"   330
```

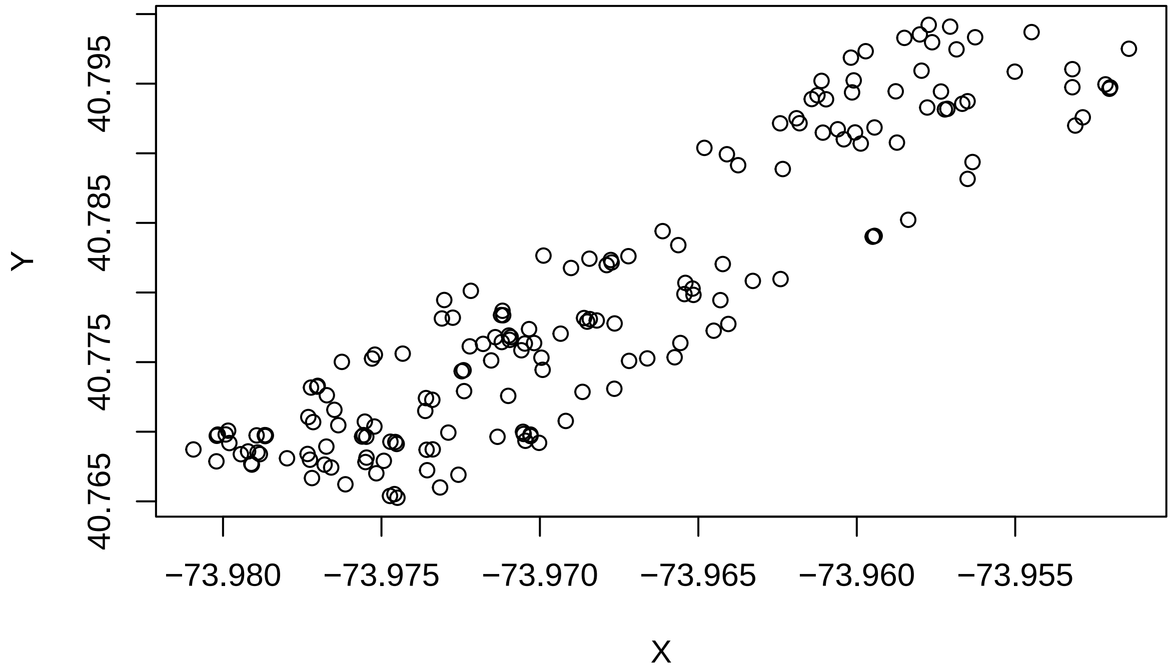
```
## Compute the percentage of adult and juveline squirrels
## who were approaching humans
df %>% filter(Age %in% c('Adult','Juvenile') & Approaches=='true') %>%
  group_by(Age) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 2 x 3
##   Age          n  freq
##   <chr>    <int> <dbl>
## 1 Adult      151 0.873
## 2 Juvenile    22 0.127
```

```
## Make a plot showing the squirrel location in the park
## (use 'X', 'Y', 'Lat/Long' is complex to use)
plot(df$X, df$Y)
```



```
## Make a plot that shows only squirrels that approach humans
df %>% filter(Approaches=='true') %>% with(plot(X, Y))
```



```
## Split date into 3 variables: month (2 first digits),
## day (2 subsequent digits), and year (4 last digits)
## demonstrate that it work correctly by printing a random
## sample of `Date` and your 3 date variables
## (but no other variables)
##
## hint: check out str_sub() function
df$month <- str_sub(df$Date,1,2)
df$day <- str_sub(df$Date,3,4)
df$year <- str_sub(df$Date,5,-1)
sighted <- df %>% sample_n(1) %>% select(Date, year, month, day)
sighted
```

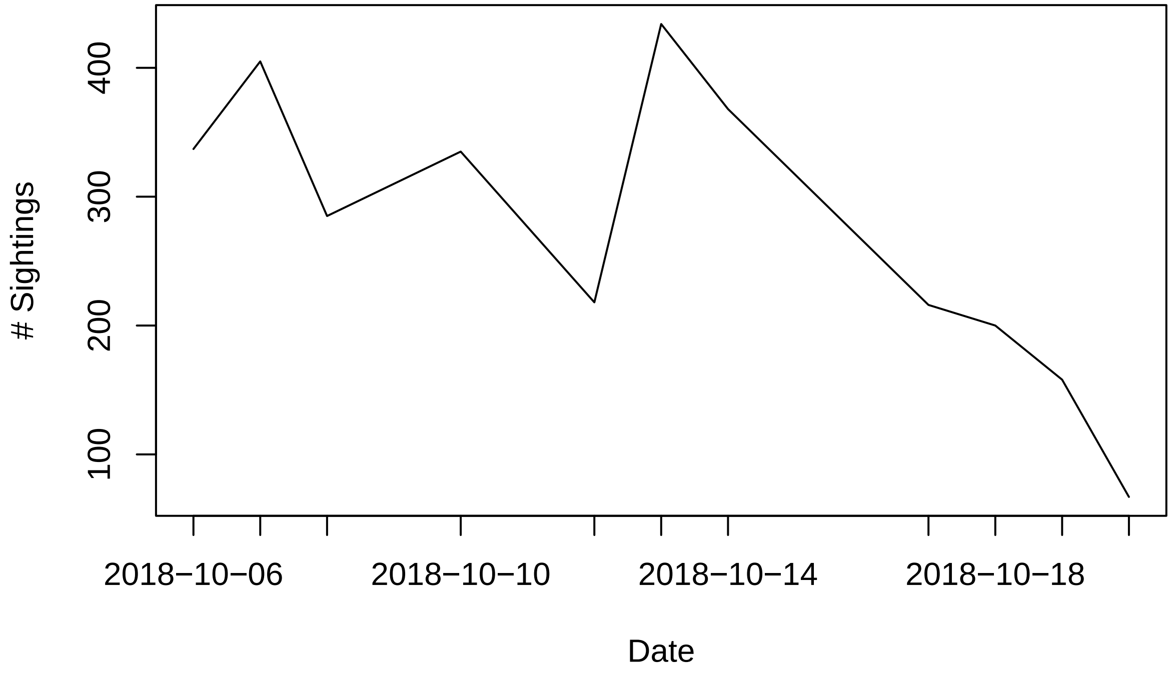
```
##       Date year month day
## 1 10142018 2018    10  14
```

```
## Compute the percentage of juvenile squirrels among those sighted
df %>% filter(Age %in% c('Adult','Juvenile') & Date==sighted$Date) %>%
  group_by(Age) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 2 x 3
##   Age          n  freq
##   <chr>    <int> <dbl>
## 1 Adult      311 0.896
## 2 Juvenile    36 0.104
```

```
## Make a line plot where you show the number of sightings by day
df$ISOdate <- as.Date(ISOdate(year = df$year, month = df$month, day = df$day)) # Convert to Date object
df_new <- df[order(df$ISOdate), ] %>% group_by(ISOdate) %>% count() # Order data

plot(df_new$ISOdate,  # Draw plot without x-axis
     df_new$n,
     type = "l",
     xaxt = "n",
     xlab = 'Date',
     ylab = '# Sightings')
axis(1,  # Add dates to x-axis
     df_new$ISOdate,
     format(df_new$ISOdate, "%Y-%m-%d"))
```


# Sightings
Date