**Scripting for Data Analysis**

**IST 652**

**Final Project**

**Katie Haugh, Mark Roberts, Sandy Spicer**

# Introduction

# The Squirrel Census, an Atlanta-based organization with a mission to count and acknowledge the Eastern gray squirrel, set out to expand their mission to the Big Apple. In 2018, an 11-day squirrel census was conducted in Central Park in New York City. The organization engaged 323 volunteers who surveyed the 843-acre park for squirrels. The 323-volunteer squirrel sighters were assigned shifts in the morning and afternoon to count squirrels. Each was assigned 1 hectare (100-by-100-meter blocks) per shift to conduct their observations for 20-25 minutes. Each countable hectare in the park – 350 in total – was counted twice (in the morning and late afternoon). The census team assigned a random number to each hectare, which signified the order to survey each hectare. The numbers were randomly assigned to remove any selective bias. The National Oceanic Atmospheric Administration’s National Weather Service station near Belvedere Castle collected the weather data from hectare 17-D. The total time sighters spent counting squirrels was nearly 368.7 hours. This project presents the survey results.

# About the Data

Two data sets are included for analysis.

*Squirrels*

The first is a .csv file containing the location, age, color, and other characteristics of 3,023 squirrels found in Central Park. The variables include:

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields** | **Description** | **Type** | **Example** |
| X | Longitude coordinate for squirrel sighting point | Number | -73.95613449 |
| Y | Latitude coordinate for squirrel sighting point | Number | 40.79408239 |
| Unique Squirrel ID | Identification tag for each squirrel sightings. The tag is comprised of "Hectare ID" + "Shift" + "Date" + "Hectare Squirrel Number." | Plain Text | 37F-PM-1014-03 |
| 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). | Plain Text | 37F |
| Shift | Value is either "AM" or "PM," to communicate whether or not the sighting session occurred in the morning or late afternoon. | Plain Text | PM |
| Date | Concatenation of the sighting session day and month. | Plain Text | 10142018 |
| Hectare Squirrel Number | Number within the chronological sequence of squirrel sightings for a discrete sighting session. | Number | 3 |
| Age | Value is either "Adult" or "Juvenile." | Plain Text | Adult |
| Primary Fur Color | Value is either "Gray," "Cinnamon" or "Black." | Plain Text | Gray |
| Highlight Fur Color | Discrete value or string values comprised of "Gray," "Cinnamon" or "Black." | Plain Text | 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. | Plain Text | + |
| Color notes | Sighters occasionally added commentary on the squirrel fur conditions. These notes are provided here. | Plain Text | Nothing selected as Primary. Gray selected as Highlights. Made executive adjustments. |
| Location | Value is either "Ground Plane" or "Above Ground." Sighters were instructed to indicate the location of where the squirrel was when first sighted. | Plain Text | Ground Plane |
| Above Ground Sighter Measurement | For squirrel sightings on the ground plane, fields were populated with a value of “FALSE.” | Plain Text | 10 |
| Specific Location | Sighters occasionally added commentary on the squirrel location. These notes are provided here. | Plain Text | Branch |
| Running | Squirrel was seen running. | Checkbox | TRUE |
| Chasing | Squirrel was seen chasing another squirrel. | Checkbox | TRUE |
| Climbing | Squirrel was seen climbing a tree or other environmental landmark. | Checkbox | FALSE |
| Eating | Squirrel was seen eating. | Checkbox | FALSE |
| Foraging | Squirrel was seen foraging for food. | Checkbox | FALSE |
| Other Activities |  | Plain Text | wrestling with mother |
| Kuks | Squirrel was heard kukking, a chirpy vocal communication used for a variety of reasons. | Checkbox | TRUE |
| Quaas | Squirrel was heard quaaing, an elongated vocal communication which can indicate the presence of a ground predator such as a dog. | Checkbox | FALSE |
| Moans | Squirrel was heard moaning, a high-pitched vocal communication which can indicate the presence of an air predator such as a hawk. | Checkbox | FALSE |
| 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. | Checkbox | TRUE |
| 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. | Checkbox | FALSE |
| Approaches | Squirrel was seen approaching human, seeking food. | Checkbox | FALSE |
| Indifferent | Squirrel was indifferent to human presence. | Checkbox | TRUE |
| Runs from | Squirrel was seen running from humans, seeing them as a threat. | Checkbox | FALSE |
| Other Interactions | Sighter notes on other types of interactions between squirrels and humans. | Plain Text | ran from dog-walker |
| Lat/Long | Latitude and longitude | Point | POINT (-73.9561344937861 40.7940823884086) |

Source: <https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw>

*Hectares*

The second data set includes location, weather, and environmental information on the 350 hectares of Central Park during the census:

|  |  |  |  |
| --- | --- | --- | --- |
| **Fields** | **Description** | **Type** | **Example** |
| 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 | Plain Text | 01A |
| Shift | Value is either "AM" or "PM," to communicate whether the sighting session occurred in the morning or late afternoon. | Plain Text | AM |
| Date | Concatenation of the sighting session month, day, and year (MMDDYYYY). | Plain Text | 10072018 |
| Anonymized Sighter | Sighter names were alphabetized then assigned a number (1-267) so that the individual abilities and characteristics of each sighter could be taken into account. | Number | 110 |
| Sighter Observed Weather Data | Each sighter was asked to describe the weather conditions before starting each sighting session. Weather values varied widely, and are in some cases inaccurate. | Plain Text | 70¬∫ F, Foggy |
| Litter | Value can be "None", "Some" or "Abundant." | Plain Text | Some |
| Litter Notes | Sighters occasionally added commentary on the amount or quality of litter. These notes are provided here. | Plain Text |  |
| Other Animal Sightings | Comma separated values indicating other animals sighted in the hectare. A list of other animals common to the park were provided; however, other animals were listed as well. | Plain Text | Humans, Pigeons |
| Hectare Conditions | Value is one of the following: “Calm”, “Moderate”, “Calm, busy” or “Busy.” The difference between “moderate” and “calm, busy” is as follows: “moderate” was either entered by the sighter, or translated from an entry that described a condition somewhere in-between busy and calm. “Calm, busy” was either entered by the sighter or translated from an entry that described discrete hectare sections as busy, but other hectare areas as separately calm. | Plain Text | Busy |
| Hectare Conditions Notes | Sighters occasionally added commentary on the hectare conditions. These notes are provided here. | Plain Text | Construction |
| Number of sighters | The number of sighters that observed the hectare for the sighting session. Values range from "1" to "3." | Number | 1 |
| Number of Squirrels | The number of squirrels sighted in the hectare during the sighting session. | Number | 4 |
| Total Time of Sighting | The sighting session duration, in minutes. | Number | 22 |

Source: <https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Hectare-Data/ej9h-v6g2>

# Data Preprocessing

## *Squirrels*

In reviewing the primary squirrels' data set, we observed several unnecessary columns and missing values. To start the cleaning process, the Unique Squirrel ID, Shift, Date, Hectare Squirrel Number, Combination of Primary and Highlight Color, Color notes, Location, Above Ground Sighter Measurements, Specific Location, Other Activities, Other Interactions, and Lat/Long columns were removed. Shift, Date, and Hectare Squirrel Number columns were duplicate columns from the hectare data frame. Combination of Primary and Highlight Color was a calculated field that combined Primary Fur Color and Highlight Fur Color with a "+". The original columns' format was ideal for our analysis, making the combination column no longer necessary. This is the same case for the Lat/Long column. The location column contained a significant number of missing values and was irrelevant for our analysis. The remaining notes columns included comments from surveyors. Most rows had missing values, and the values that were present were inconsistent in context.

Next, we addressed missing values in the remaining columns. For the age and primary fur color columns, if a value was missing, it was replaced with "Unknown." For the highlight fur color column, we exchanged blanks with "None." The reason for the difference is because some squirrels were only one color.

## *Hectares*

The hectares data set also contained a few extraneous columns. Anonymized Sighter, which identifies who surveyed the hectare, was removed along with three columns pertaining to the litter and condition of the hectare. Since these columns were primarily blank, we decided to remove them. The date column was then correctly formatted.

Additionally, we observed two columns that needed to be normalized. The Sighter Observed Weather Data column contained the temperature and weather conditions (ex. 70°F, Foggy.) First, we separated the temperature from the condition, and then the condition was categorized into one of four main conditions: cloudy, sunny, rainy, and other. The other column requiring normalization was the Other Animal Sightings column. This column contained a list of animals in each row. Four more categories, Humans, Dogs, Birds, and Rodents, were created to classify each animal. Each category became their own column accepting a binary value.

# Business Questions

## *Question 1*

*What are the primary visual characteristics of the squirrels? What color combinations were most popular? Were squirrels of similar markings found in a particular area of the park?*

To evaluate this question, the Primary Fur Color and Highlight Fur Color columns in the squirrel dataset will be used to look for counts of each combination. Additionally, combining that information with the Hectare, X, and Y location columns, we can evaluate populated areas of the park by fur colors.

## *Question 2*

*Is there a section of the park that tends to have more squirrels? What are the characteristics of those hectares? Were other animals present?*

Using the hectare dataset, the number of squirrels and hectare columns were used to determine the most populated part of the park. Then using the Other Animal Sighting column, those areas were evaluated for animal activity.

## *Question 3*

*Is there a relationship between the number of squirrels seen and the time of day? Does the weather at the time have any influence?*

To summarize the squirrel sightings by time of day and weather, the hectare dataset was broken out by shift and weather condition. The temperature was then categorized in buckets of 10 degrees and summarized.

## *Question 4*

*What verbal noises were most popular? What behaviors are most common? How do these distribute around the park?*

Using the squirrel dataset, we calculated the counts of types of sounds ("Kuks," "Quaas," and "Moans") as well as counts of squirrel behaviors (tail flagging/twitching, approaching/running from and indifference). The distribution of sounds is mapped on a Central Park map based on coordinate data.

# Program Description

The program starts out by reading in the squirrel csv into a pandas dataframe and cleaning the dataframe via the outline in the cleaning section above. The same process then occurs for the hectare dataset. Some general descriptive statistics are performed to evaluate the distribution of the data and any correlations.

## *Question 1*

## The program creates a column to classify each squirrel as either in the lower, middle, or upper part of the park. A summary table is then created, grouping the squirrels by primary fur color. A plot of the distribution of the primary fur colors is then produced. The same summary table and bar chart is created for the highlight fur colors. Next, the program combines primary fur color and highlight fur colors into one summary table with 22 combinations displaying the count of squirrels for each combination and the percent of squirrels with each marking combination. A bar chart of this summary table is outputted. The squirrels are then mapped and colored by their primary fur color. A second map is created, removing the gray squirrels to focus on the clusters of the other primary fur colors. Next, a summary table of primary fur color by park area is created, along with an associated bar chart.

## *Question 2*

Next, the program creates a summary table of the number of squirrels in each park area. A bar chart of this data is then outputted. The same park area column on the squirrel dataset is then added to the hectare dataset. Then a summary table was created, counting the number of human, dog, bird, and rodent sightings for each section of the park. The program then melts the table into a long table with a row for each area of the park, and then a bar chart was produced.

## *Question 3*

The program then creates a dataframe summarizing the number of squirrels by time of day. A bar chart of this data is generated. Next, the program creates a dataframe summarizing the number of squirrels by weather condition and creates a bar chart. The program then combines the time of day and weather information by creating a dataframe to summarize the number of squirrels by weather condition and shift. Another bar chart is created. Next, five temperature buckets are created (40s, 50s, 60s, 70s, and 80s) and added to a temp\_range dataframe. The program then adds this as a column to the hectare dataframe. Finally, a dataframe is created to summarize the number of squirrels by temperature range and generates a bar chart.

## *Question 4*

Finally, the program creates a categorical column for the three "verbalizations" from the survey: "Kuks", "Quaas" and "Moans", as it were; subsequently, it provides a count and percentage of sounds by categorical name. Similarly, the program counts the number of each squirrel behavior observed by sighters, including "Tail flags", "Tail twitches," "Approaches," "Indifferent," and "Runs from." A further visualization shows the distribution of noises around the park, utilizing coordinate data points provided by the survey.

# Analysis Results

## *Question 1*

*What are the primary visual characteristics of the squirrels? What color combinations were most popular? Were squirrels of similar markings found in a particular area of the park?*

To evaluate the visual characteristics of the squirrels, we first summarized the squirrels by their primary fur color. There were three main colors, black, cinnamon, and gray, and some that had NAs. We categorized those NAs as unknown and created a separate category called unknown. Of those colors, 82% (2,473) of the central park squirrels were gray, 13% (392) were cinnamon, and only a small margin had primarily black fur.

Chart, waterfall chart

Description automatically generated

Next, we summarized the squirrels by their highlight fur color. There were four main colors, black, cinnamon, gray, and white. Since some had NAs, we categorized those NAs as none, creating a separate category. Of those colors, 36% (1,086) of the Central Park squirrels had no highlight color, 35% (1.076) had cinnamon markings, and 31% (954) had white markings. Black and/or gray highlights were only seen in 10% of squirrels.

Chart, bar chart

Description automatically generated

To combine the primary and highlight fur markings, we created a summary table of all the combinations of colors. There were 22 color combinations: six black, seven cinnamon, eight gray, and one unknown. Most black squirrels had no highlight fur colors, most cinnamon squirrels had gray highlights, and most gray squirrels had no highlights. Gray squirrels were the only ones seen with up to three highlight colors.

Chart

Description automatically generated

To evaluate if squirrels of similar markings were found in a particular part of the park, we first divided the hectares into lower, middle, and upper sections. This made it easier to summarize the data. For the most part, a majority of squirrels from each primary fur color were found in the lower part of the park.

Chart, waterfall chart

Description automatically generated

We then mapped all the squirrels to visually see where the squirrels were found. The gray squirrels were fairly distributed across the entire park, but the black squirrels were clustered in the park's upper and lower areas. Additionally, we can see that the water reservoir in the center of Central Park contained no squirrels, which may have skewed the data in this section of the analysis.

Map

Description automatically generated

## *Question 2*

*Is there a section of the park that tends to have more squirrels? What are the characteristics of those hectares? How busy were those areas? Were other animals present?*

Here we summarized the number of squirrels in each area of the park. Each section had the same number of hectares. The middle area of the park was largely covered by water, potentially being the cause for the lower number of squirrel sightings. The upper and lower sections of the part had comparable amounts of water coverage but had a 10% difference in the number of squirrels.

Application

Description automatically generated with medium confidence

After looking at the area, we summarized the animal sightings by area of the park. Overall, the animal sightings were not linked to a particular region of the part except for rodents. Most of the hectares with rodents were seen in the upper portion of the park.

Chart

Description automatically generated

## *Question 3*

*Is there a relationship between the number of squirrels seen and the time of day? Does the weather at the time have any influence?*

First, we summarized the number of squirrels observed by shift. The sightings occurred either in the morning (AM) or afternoon/evening (PM). As shown in the table and bar plot below, nearly 22% more squirrels were observed in the PM hours than the AM hours.

Chart

Description automatically generated

Next, we summarized the squirrel sightings by weather condition. The four weather condition categories used were: cloudy, other, rainy, or sunny. The majority of squirrels were observed in cloudy conditions or other weather conditions. Squirrels were moderately observed in sunny weather and very rarely observed in rainy weather.

Chart, bar chart

Description automatically generated

We also broke down the weather conditions by shift to see if certain weather conditions had more sightings in the AM or PM. As shown in the table and bar plot below, most squirrels were observed in the PM across all weather conditions except in the rain.

Chart, bar chart

Description automatically generated

After diving into weather conditions, we explored the number of squirrels observed across different temperature ranges. As shown in the table and bar plot below, most squirrels were observed in temperatures ranging in the 50s and 60s. There were fewer squirrel sightings on colder days in the 40s and hotter days in the 70s & 80s.

Chart, bar chart

Description automatically generated

## *Question 4*

*What verbal noises were most popular? What behaviors are most common? How do these distribute around the park?*

With 87 observations, "Kuks" represent 63% of the total vocalizations, and nearly one-third of vocalizations are “Quaas”. "Moan" barely registers on the tally with only three observations, at just 2% of occurrences. Whatever meaning is expressed by "Moans" is undoubtedly a squirrel event that may have eluded the sighters.

Chart, bar chart

Description automatically generated

Behaviors show a broader distribution, with "Indifferent" squirrels on top at 43, while "Runs from" is a close second at 39. "Tail flags" and "Tail twitches are at 18 and 17, respectively, with "Approaches" at the bottom with four counts.

Chart, bar chart

Description automatically generated

There does not appear to be a pattern in the distribution of verbalization types throughout the park. Whatever the cause of one sound or another does not appear to be locally specific.

Diagram

Description automatically generated with low confidence

# Conclusions

*Squirrels*

Overall, there were 3,023 squirrels sighted in New York City’s Central Park during the 2018 squirrel census. They were of three primary fur colors, and four highlighted fur colors. There were 22 combinations of colors. Most of the squirrels sighted were gray and either had no highlights or cinnamon highlights. Central Park squirrels made two possible noises and had varying behavior towards humans. They are most likely to "Kuk" while being indifferent or perhaps running away from humans. Also, squirrel vocalizations occurred throughout the park in no discernable pattern. We didn't observe any upper or lower patois groupings.

*Hectares*

The squirrels were observed at two different times of the day over 11 days. Most squirrels were recorded in the PM, during cloudy weather conditions, and temperatures in the 50s-60s. There were less squirrel sightings in the AM, rainy conditions, and in extreme cold/warm. Additionally, most squirrels were sighted in the lower area of the park, which may have been due to less observable areas in the middle and upper regions because of a large water reservoir bordering those areas. Humans were the most observed “animal” in the park and were evenly sighted throughout the park.

# Groupwork

The team worked together on the overall structure of the program and report. Each team member took the responsibilities laid out below:

* Squirrels Data Cleaning – Mark Roberts
* Hectare Data Cleaning – Sandy Spicer
* Business Question 1 – Katie Haugh
* Business Question 2 – Katie Haugh
* Business Question 3 – Sandy Spicer
* Business Question 4 – Mark Roberts