WE, THE NEW YORK SQUIRRELS

A data analysis project

UNIVERSITY OF WESTERN AUSTRALIA
DATA ANALYTICS BOOTCAMP - 2023

“Have you ever wondered how many squirrels are in New York City parks, where they are located, and what they are up to?”

This data analysis project aims to tell this story, and specifically answer the following questions:

- How does the squirrel frequency and range of activities and behaviours compare between autumn and spring?
- How does the squirrel population diversity (colours, age, location) compare between autumn and spring?
- How do squirrels, in a park setting, interact with humans and other wildlife?

Project Theme

ANALYSTS

Alyssa Hondrade, Lakna Premasinghe, Adam Mayer and Dominique Spencer

SCOPE

The project will create an interactive visualisation of a squirrel's day-to-day life in New York City; inspired by and using data from *The Squirrel Census*, an annual data collection initiative run by scientists and volunteers, that aims to tell a story about the local Eastern Grey squirrel population.

The project is based on the 2018 and 2020 Squirrel Census, where the data from 2018 was collected over a two-week period and the data for 2020 was collected on a single day (01 March 2020) - this was simplified for the project as *autumn* and *spring*.

The project singles in on a specific animal but can be scaled to any wildlife and its interactions/ behaviours in similar metropolitan areas.

The project will focus on the following data attributes:

Location (latitude and longitude)
Squirrels identified using unique IDs
Primary and Highlight Fur Colour
Activities (running, chasing, climbing, eating, foraging, etc)
Interactions with humans (approaches, indifferent, runs from)
The date of the sighting

Sources (acorns)

THE SQUIRREL CENSUS
DEPT. OF DATA — NYC

The Squirrel Census

"The Squirrel Census is a multimedia science, design, art, and storytelling project focusing on the Eastern gray (Sciurus carolinensis). We count squirrels and present our findings to the public in the form of comprehensive maps, data visualizations, 45-RPM records, anthropomorphic storytelling, live presentations, semi-true videos, and other fun."

2020 dataset. The raw dataset contains two separate csv files; a squirrel dataset and a park dataset.
The squirrel dataset includes 433 entries, and the park dataset contains 25 entries.
The cleaned, merged version for this project includes a total of 192 entries.

NYC OpenData

NYC OpenData

"NYC Open Data is managed by the Open Data Team at the NYC Office of Technology and Innovation (OTI). The team works with City agencies to identify and make data available, coordinate platform operations and improvements, and promote the use of Open Data both within government and throughout NYC. We believe that every New Yorker can benefit from Open Data, and Open Data can benefit from every New Yorker."

2018 dataset. The raw dataset contains one csv file containing 3023 entries.
The cleaned version for this project includes 1,933 entries.

THE SQUIRREL CENSUS

DEPT. OF DATA — NYC

C-SQRL CALL NOW! 1-833-NYC-SQ

Definitions and limitations

Seasons

Spring: data collected in October, from the 2018 dataset

Autumn: data collected in March, from the 2020 dataset

Squirrels

Squirrels were not counted more than once, the scientific advisor onsite.

Parks

2018 data was collected in Central Park only.

2020 data was collected across numerous NYC parks.

Timeframe

2018 data was collected over a two week time period.

2020 data was collected on one day only.

Coding Approach

A step-by-step approach was taken when pseudo-coding, analysing, debugging and building the code for this project.

A number of points to note include:

Track : Dashboard page with multiple charts that update from the same data.

Database : SQLite

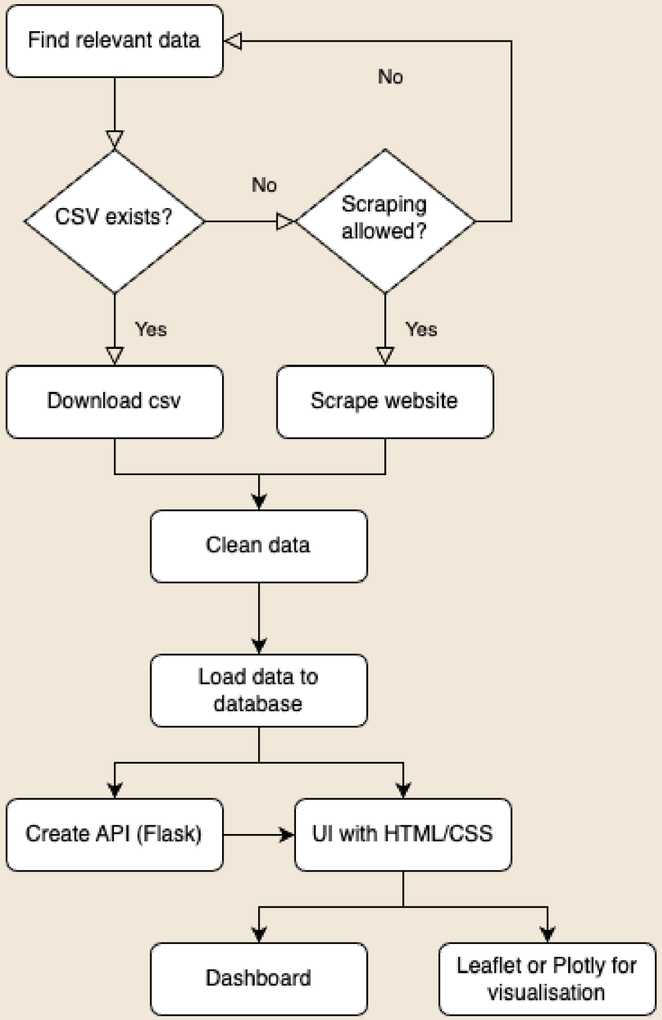
Additional JS Library : Chroma.js (colour scale) and Lodash (utility library)

Jupyter Notebook was the primary testing-pad for blocks of code used throughout.

More fluent analysts built the framework for each step, which were then contributed to and filled in by others.

Frequent meetings were held to float ideas, discuss next steps and walk through changes that had been made to documents.

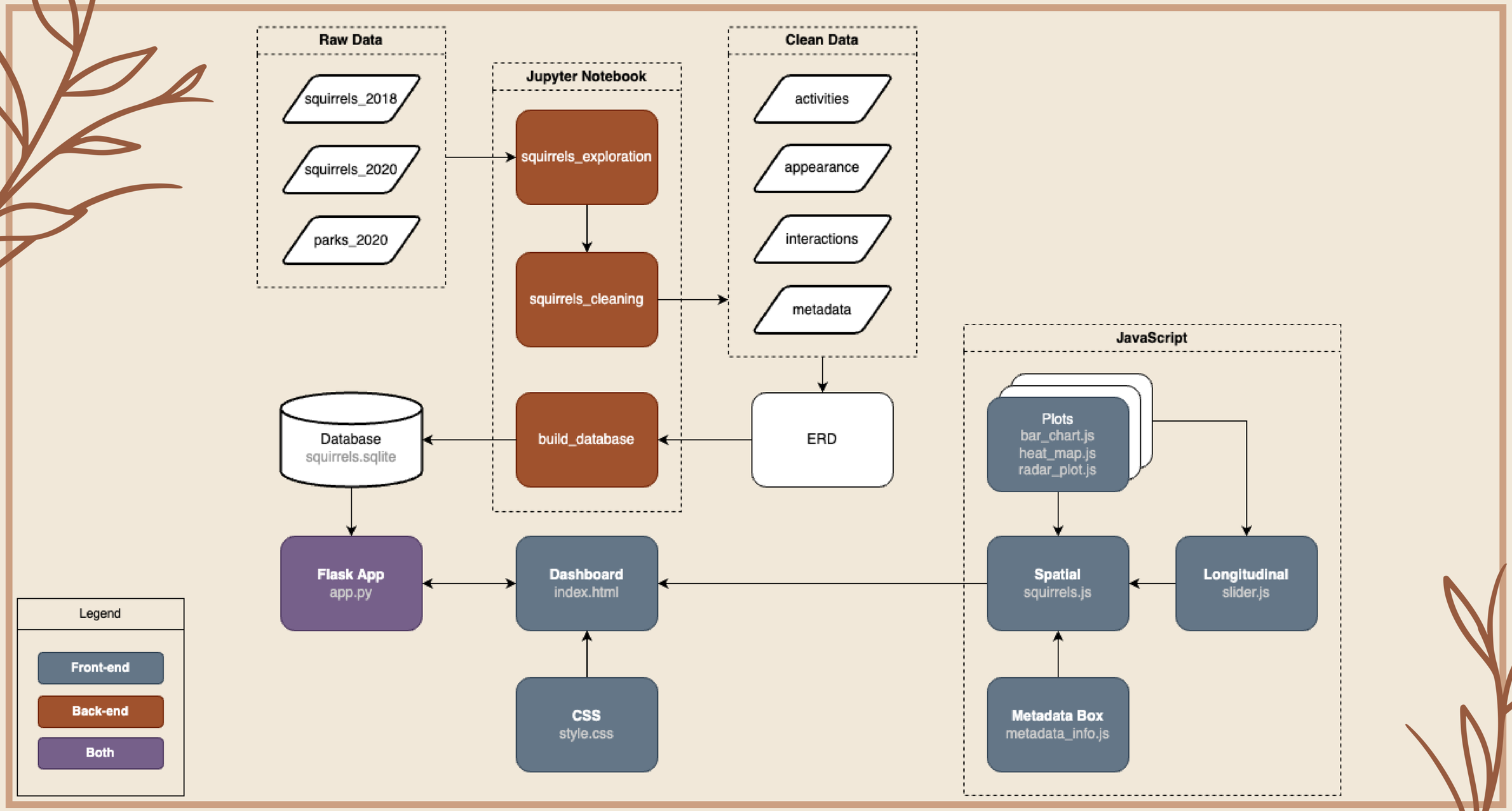
Project Workflow



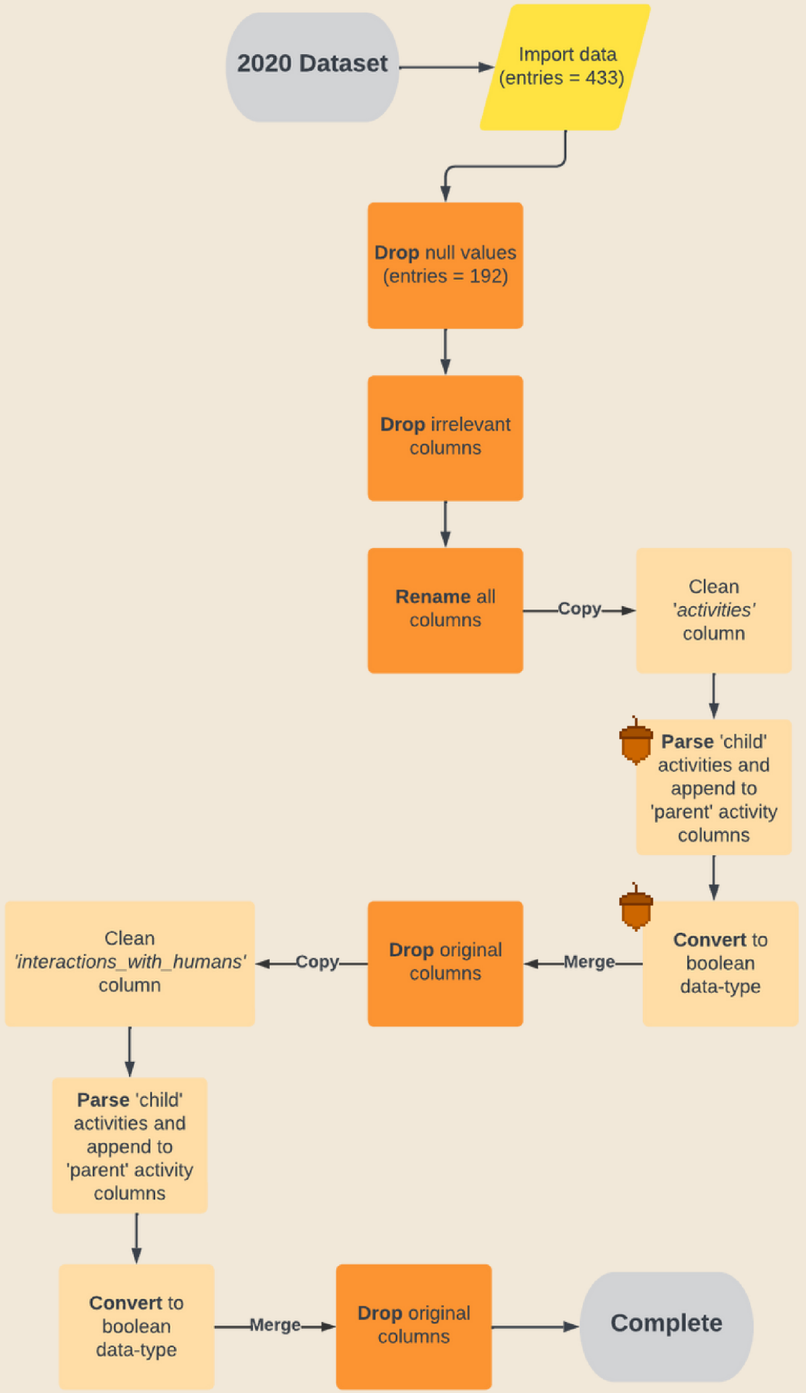
Decision log

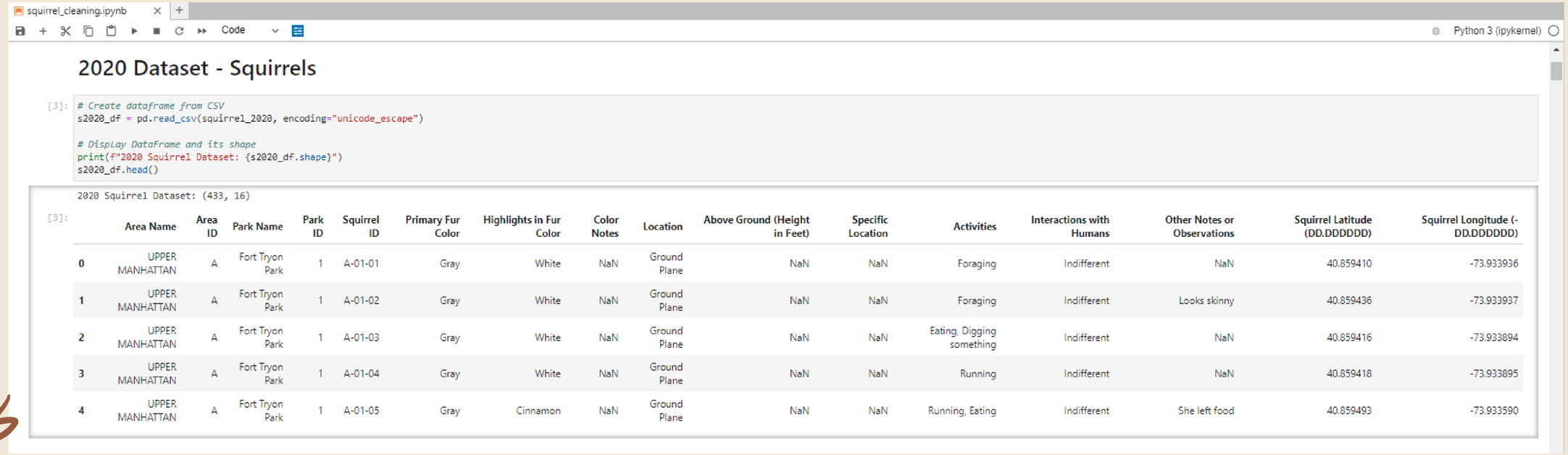
decision_log.md

```
# Decision Log
The purpose of this document is to log all decisions to a central location accessible to all team members.

## Data Cleaning
1. (02 Nov) "Other Notes or Observations / Interactions"
- Description:
    - 2020 Dataset "Other Notes or Observations": 192 >> 96 rows, with 76 unique
    - 2018 Dataset "Other Interactions": 1937 >> 170, with 152 unique
    - Can include NaN values, use as an extra point of interest?
- Decision: No need to use dropna(), retain rows with null values since not a primary attribute.

2. (02 Nov) "Location / Above Ground / Specific Location"
- Description: If retained, will require extra work to parse.
- Decision: Since not a primary attribute and resource-limited, drop columns.

3. Columns Cleaning
- Remove columns (2020):
    - 'Area Name', irrelevant.
    - 'Area ID, irrelevant.
    - 'Color Notes', irrelevant and too few data points.
    - 'Location', irrelevant and would take too much resources to clean the data. ***
    - 'Above Ground (...)', irrelevant and would take too much resources to clean the data. ***
    - 'Specific Location', irrelevant and would take too much resources to clean the data.
- Remove columns (2018):
    - 'Hectare', inconsistent with 2020 dataset.
    - 'Shift', inconsistent with 2020 dataset.
    - 'Hectare Squirrel Number', inconsistent with 2020 dataset.
    - 'Age', inconsistent with 2020 dataset.
    - 'Date', inconsistent with 2020 dataset, right??? We know 2018 data was collected in October [Autumn] and 2020 data was collected in March [Spring]
    - 'Combination of Primary and Highlight Color', inconsistent with 2020 dataset.
    - 'Color notes', inconsistent with 2020 dataset.
    - 'Location', irrelevant and would take too much resources to clean the data. ***
    - 'Above Ground Sighter Measurement', irrelevant and would take too much resources to clean the data. ***
    - 'Specific Location', inconsistent with 2020 dataset.
    - 'Lat/Long', delete due to double up.

4. Finalise column names
    - park_name                 /  - park_name [Central Park only - can create new column to match 2020 dataset column]
    - park_ID                   /  - park_ID [DOESN'T EXIST IN 2018 DATASET - or maybe we keep this as an ID might be easier to map?]
    - squirrel_ID               /  - 'Unique Squirrel ID' (change to squirrel_ID)
    - primary_fur_color         /  - primary_fur_color
    - highlights_in_fur_color   /  - 'Highlight Fur Color' (change to highlights_in_fur_color)
    - foraging
    - climbing
    - eating
    - running
    - chasing
    - shouting                  /  - 'Kuks' + 'Quaas' + 'Moans' (change to shouting)
    - sitting
    - digging
    - other_activities          /  - 'Other Activities' + 'Tail flags' + 'Tail twitches' (changed to other_activities)
    - interactions_with_humans  /  - 'Approaches' + 'Indifferent' + 'Runs from' + 'Other Interactions' (changed to interactions_with_humans) [MAYBE 'OTHER SPECIES'?]
    - other_observations        /  - other_observations [DOESN'T EXIST IN 2018 DATASET]
    - squirrel_latitude         /  - 'Y' (change to squirrel_latitude)
    - squirrel_longitude        /  - 'X' (change to squirrel_longitude)

5. The "highlights" column needed to be parsed to access the different highlight colours. This was discovered after merging the two datasets, and processed accordingly.

6. Needed to change datatypes constantly, since "True" and "False" was not being recognised as a boolean.

```


Raw Data
squirrels_2018
squirrels_2020
parks_2020
Jupyter Notebook
squirrels_exploration
squirrels_cleaning
build_database
Clean Data
activities
appearance
interactions
metadata
ERD
Database
squirrels.sqlite
Flask App
app.py
Dashboard
index.html
CSS
style.css
JavaScript
Plots
bar_chart.js
heat_map.js
radar_plot.js
Spatial
squirrels.js
Longitudinal
slider.js
Metadata Box
metadata_info.js
Legend
Front-end
Back-end
Both

Data Wrangling Techniques


2020 Dataset
Import data (entries = 433)
Drop null values (entries = 192)
Drop irrelevant columns
Rename all columns
Copy
Clean 'activities' column
Parse 'child' activities and append to 'parent' activity columns
Convert to boolean data-type
Merge
Drop original columns
Copy
Clean 'interactions_with_humans' column
Parse 'child' activities and append to 'parent' activity columns
Convert to boolean data-type
Merge
Drop original columns
Complete



	Area Name	Area ID	Park Name	Park ID	Squirrel ID	Primary Fur Color	Highlights in Fur Color	Color Notes	Location	Above Ground (Height in Feet)	Specific Location	Activities	Interactions with Humans	Other Notes or Observations	Squirrel Latitude (DD.DDDDDD)	Squirrel Longitude (-DD.DDDDDD)
0	UPPER MANHATTAN	A	Fort Tryon Park	1	A-01-01	Gray	White	NaN	Ground Plane	NaN	NaN	Foraging	Indifferent	NaN	40.859410	-73.933936
1	UPPER MANHATTAN	A	Fort Tryon Park	1	A-01-02	Gray	White	NaN	Ground Plane	NaN	NaN	Foraging	Indifferent	Looks skinny	40.859436	-73.933937
2	UPPER MANHATTAN	A	Fort Tryon Park	1	A-01-03	Gray	White	NaN	Ground Plane	NaN	NaN	Eating, Digging something	Indifferent	NaN	40.859416	-73.933894
3	UPPER MANHATTAN	A	Fort Tryon Park	1	A-01-04	Gray	White	NaN	Ground Plane	NaN	NaN	Running	Indifferent	NaN	40.859418	-73.933895
4	UPPER MANHATTAN	A	Fort Tryon Park	1	A-01-05	Gray	Cinnamon	NaN	Ground Plane	NaN	NaN	Running, Eating	Indifferent	She left food	40.859493	-73.933590

Parse individual words from ‘activities’ column in a **for loop** and add to respective ‘parent’ column

squirrel_cleaning.ipynb

Code

[12]:
```
pattern = 'eating \((.*?)\)'
sitting = ["hanging out", "hanging", "chilling", "sitting in short tree", "sticking out of a tree", "very carefully watching a cat", "watching #2", "posing"]
shouting = ["vocalization at us", "defending tree"]
other_activities = ["self-cleaning", "sleeping", "sleeping (dead?)", "battery"]

for row in activities_df['activity_list']:
    for idx, word in enumerate(row):
        row[idx] = word.lower()
        word = word.lower()

        # Check for "eating (...)"
        bracket = re.findall(pattern, word)
        if (len(bracket) > 0):
            row[idx] = "eating"

        # Check for "digging "
        if (word.startswith("digging ") or word == "burying"):
            row[idx] = "digging"

        # Check for words in the shouting list
        if (word in shouting):
            row[idx] = "shouting"

        # Convert to "foraging"
        if (word == "nesting/gathering leaves"):
            row[idx] = "foraging"

        # Check for words in the sitting list
        if (word in sitting):
            row[idx] = "sitting"

        # Convert to "climbing"
        if (word == "jumping"):
            row[idx] = "climbing"

        # Account for other_activities
        if (word in other_activities):
            row[idx] = "other_activities"
```

[13]:
```
# Check the 'activities' distribution
flat_list = [value for row in activities_df['activity_list'] for value in row]
flat_df = pd.DataFrame(flat_list)
flat_df.value_counts()
```

[13]:
```
foraging          72
climbing          52
eating            48
running           37
chasing           20
shouting          17
sitting           14
```

Parsing each ‘activity’ into its own boolean-datatype columns, then **drop** the original column

squirrel_cleaning.ipynb

Code

[196]:
```
### Create the boolean activity columns

# Create a list of the reduced activity names
column_names = list(set([word.lower() for word in flat_list]))

# Create new columns for each activity
for col_name in column_names:
    activities_df[col_name] = pd.Series(dtype=bool)

# Loop through the 'activity_list' and populate the boolean columns
for idx, row in activities_df.iterrows():
    for word in row['activity_list']:
        for activity in column_names:
            if (row[activity] == True):
                continue
            if (word == activity):
                row[activity] = True
            else:
                row[activity] = False

# Display the DataFrame
activities_df.head()
```

[196]:

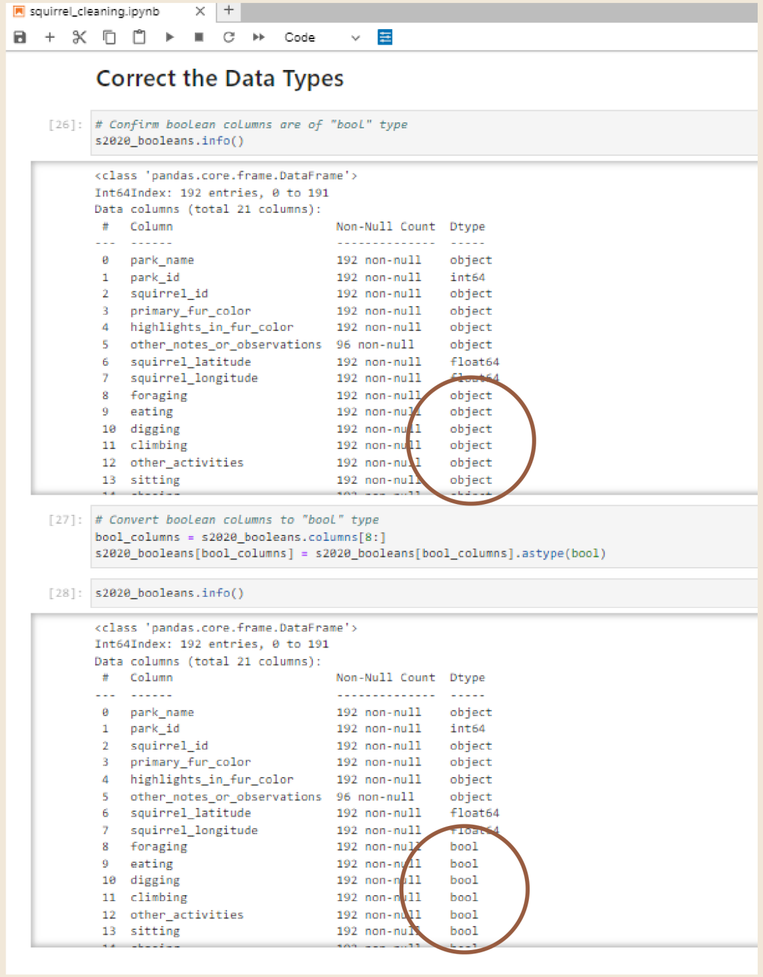
	squirrel_id	activities	activity_list	shouting	chasing	eating	climbing	sitting	other_activities	running	digging	foraging
0	A-01-01	Foraging	[foraging]	False	False	False	False	False	False	False	False	True
1	A-01-02	Foraging	[foraging]	False	False	False	False	False	False	False	False	True
2	A-01-03	Eating, Digging something	[eating, digging]	False	False	True	False	False	False	False	True	False
3	A-01-04	Running	[running]	False	False	False	False	False	False	True	False	False
4	A-01-05	Running, Eating	[running, eating]	False	False	True	False	False	False	True	False	False

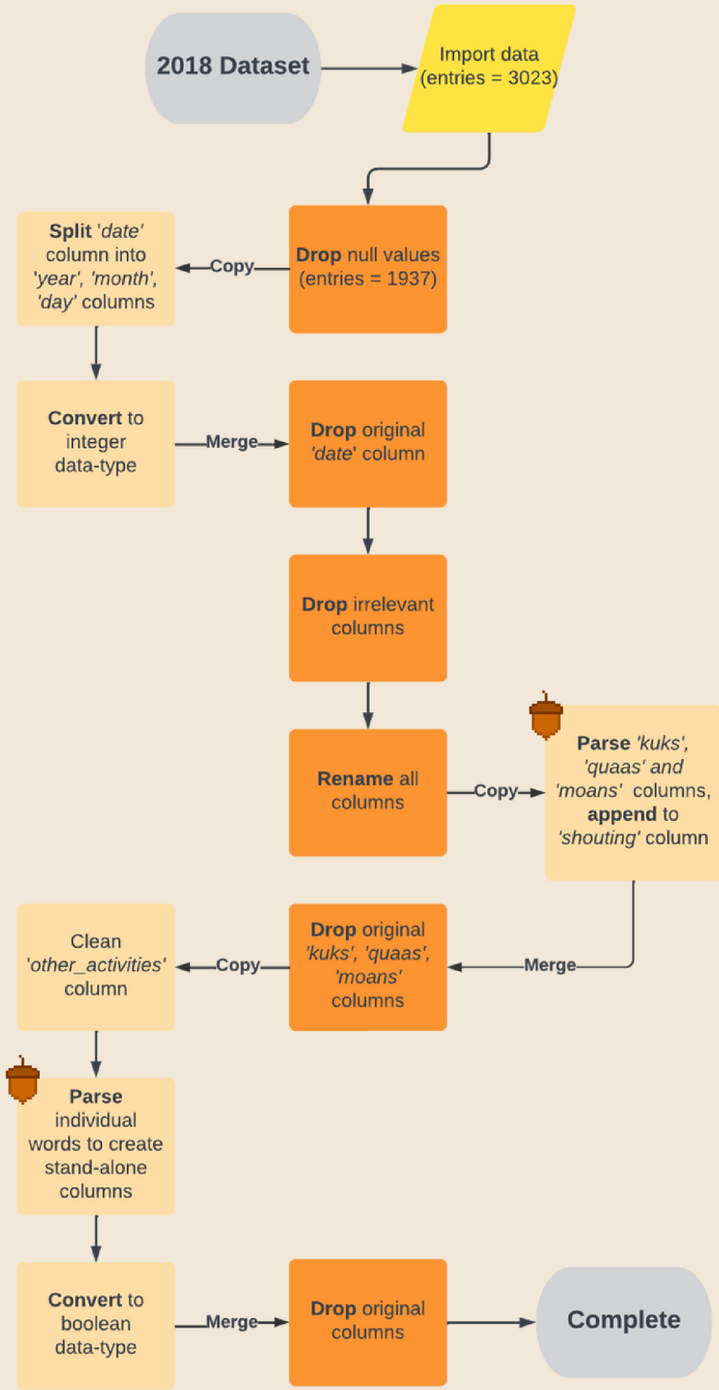
[15]:
```
# Drop the 'activities' and 'activity_list' columns
activities_df = activities_df.drop(columns=['activities', 'activity_list'])
activities_df.head()
```

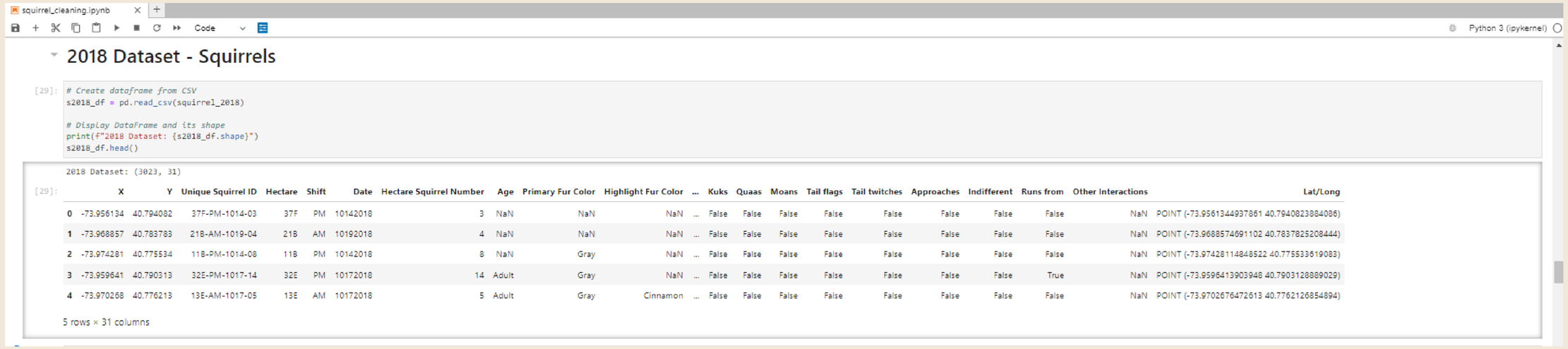
[15]:

	squirrel_id	shouting	chasing	eating	climbing	sitting	other_activities	running	digging	foraging
0	A-01-01	False	False	False	False	False	False	False	False	True
1	A-01-02	False	False	False	False	False	False	False	False	True
2	A-01-03	False	False	True	False	False	False	False	True	False
3	A-01-04	False	False	False	False	False	False	True	False	False

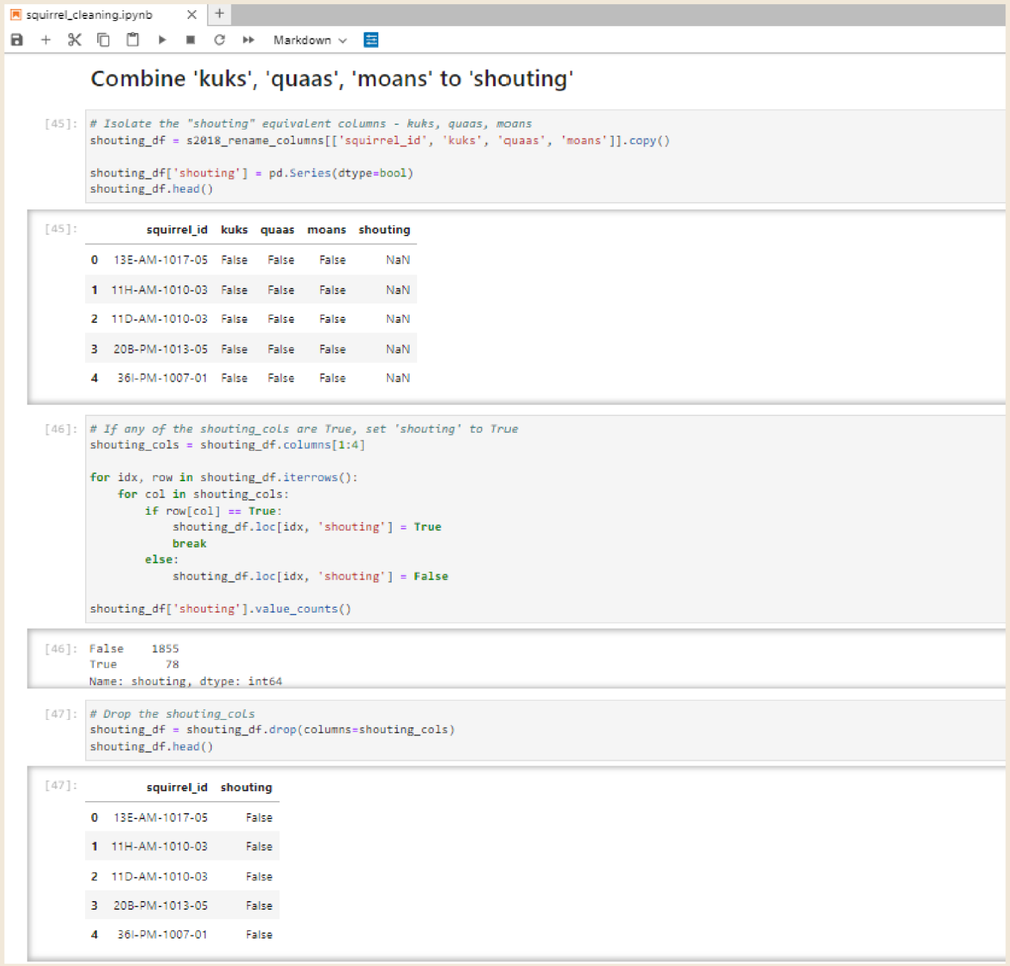
New ‘activities’ and ‘interactions’ columns were **merged** back into the main dataframe, and data-types were corrected from ‘object’ to ‘boolean’




2018 Dataset
Import data (entries = 3023)
Drop null values (entries = 1937)
Copy
Split 'date' column into 'year', 'month', 'day' columns
Drop original 'date' column
Merge
Convert to integer data-type
Drop irrelevant columns
Rename all columns
Copy
Parse 'kuks', 'quaas' and 'moans' columns, append to 'shouting' column
Drop original 'kuks', 'quaas', 'moans' columns
Merge
Clean 'other_activities' column
Copy
Parse individual words to create stand-alone columns
Drop original columns
Merge
Convert to boolean data-type
Complete


squirrel_cleaning.ipynb
Code
Python 3 (ipykernel)
2018 Dataset - Squirrels
[29]:
# Create dataframe from CSV
s2018_df = pd.read_csv(squirrel_2018)
# Display DataFrame and its shape
print(f"2018 Dataset: {s2018_df.shape}")
s2018_df.head()
2018 Dataset: (3023, 31)
[29]:
X Y Unique Squirrel ID Hectare Shift Date Hectare Squirrel Number Age Primary Fur Color Highlight Fur Color ... Kuks Quaas Moans Tail flags Tail twitches Approaches Indifferent Runs from Other Interactions Lat/Long
0 -73.956134 40.794082 37F-PM-1014-03 37F PM 10142018 3 NaN NaN NaN ... False False False False False False False False NaN POINT (-73.9561344937861 40.7940823884086)
1 -73.968857 40.783783 21B-AM-1019-04 21B AM 10192018 4 NaN NaN NaN ... False False False False False False False False NaN POINT (-73.9688574691102 40.7837825208444)
2 -73.974281 40.775534 11B-PM-1014-08 11B PM 10142018 8 NaN Gray NaN ... False False False False False False False False NaN POINT (-73.97428114848522 40.775533619083)
3 -73.959641 40.790313 32E-PM-1017-14 32E PM 10172018 14 Adult Gray NaN ... False False False False False False False True NaN POINT (-73.9596413903948 40.7903128889029)
4 -73.970268 40.776213 13E-AM-1017-05 13E AM 10172018 5 Adult Gray Cinnamon ... False False False False False False False False NaN POINT (-73.9702676472613 40.7762126854894)
5 rows × 31 columns

Like the ‘activities’ and ‘interactions’ columns from the 2020 dataset, some columns needed data to be separated and **parsed** into independent columns



squirrel_cleaning.ipynb

Markdown

Combine 'kuks', 'quaas', 'moans' to 'shouting'

[45]:
```
# Isolate the "shouting" equivalent columns - kuks, quaas, moans
shouting_df = s2018_rename_columns[['squirrel_id', 'kuks', 'quaas', 'moans']].copy()

shouting_df['shouting'] = pd.Series(dtype=bool)
shouting_df.head()
```

[45]:

	squirrel_id	kuks	quaas	moans	shouting
0	13E-AM-1017-05	False	False	False	NaN
1	11H-AM-1010-03	False	False	False	NaN
2	11D-AM-1010-03	False	False	False	NaN
3	20B-PM-1013-05	False	False	False	NaN
4	36I-PM-1007-01	False	False	False	NaN

[46]:
```
# If any of the shouting_cols are True, set 'shouting' to True
shouting_cols = shouting_df.columns[1:4]

for idx, row in shouting_df.iterrows():
    for col in shouting_cols:
        if row[col] == True:
            shouting_df.loc[idx, 'shouting'] = True
            break
        else:
            shouting_df.loc[idx, 'shouting'] = False

shouting_df['shouting'].value_counts()
```

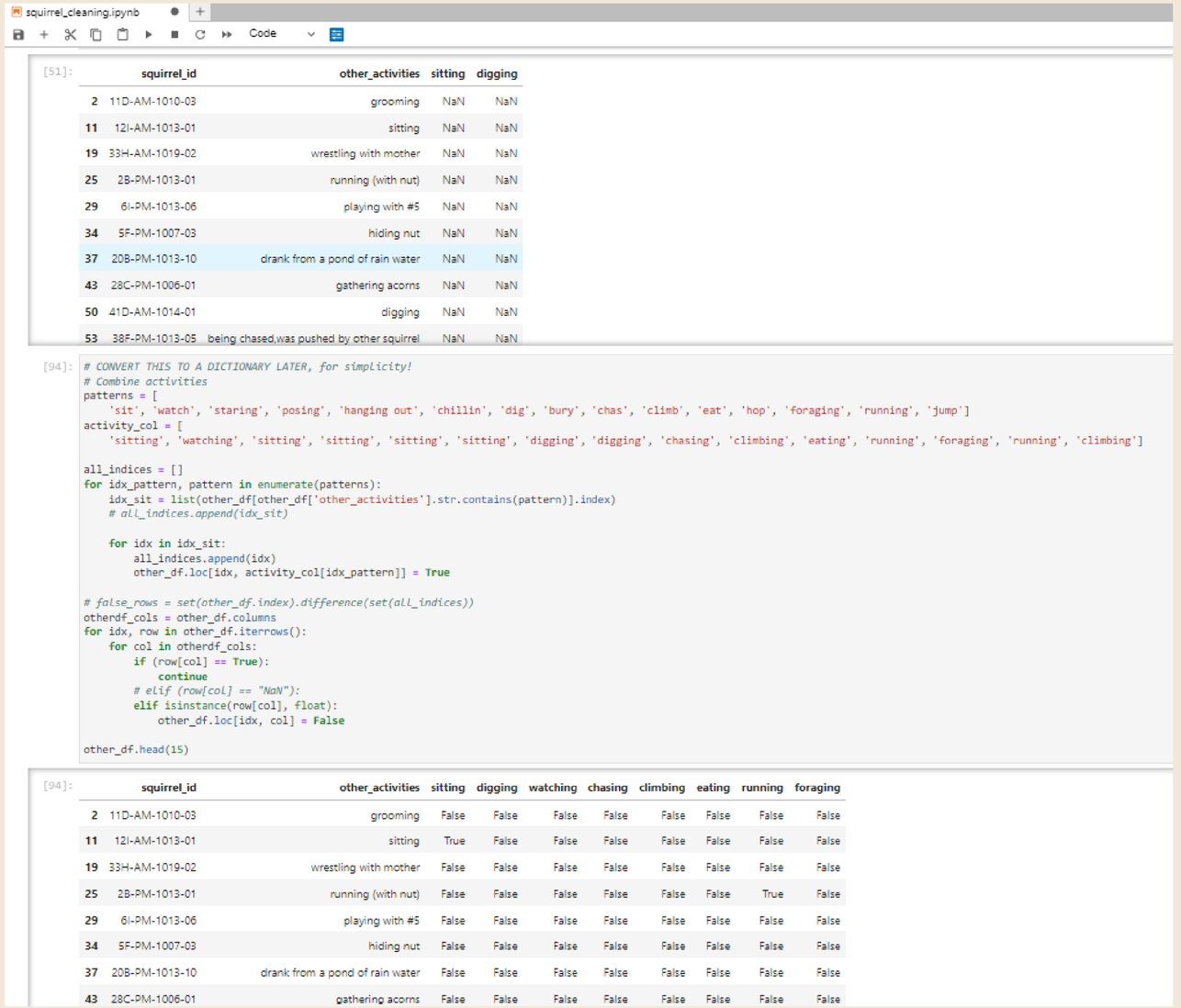
[46]:
```
False    1855
True       78
Name: shouting, dtype: int64
```

[47]:
```
# Drop the shouting_cols
shouting_df = shouting_df.drop(columns=shouting_cols)
shouting_df.head()
```

[47]:

	squirrel_id	shouting
0	13E-AM-1017-05	False
1	11H-AM-1010-03	False
2	11D-AM-1010-03	False
3	20B-PM-1013-05	False
4	36I-PM-1007-01	False

Parse individual words from ‘other_activities’ column and add to new ‘parent’ columns



squirrel_cleaning.ipynb

Code

[51]:

	squirrel_id	other_activities	sitting	digging
2	11D-AM-1010-03	grooming	NaN	NaN
11	12I-AM-1013-01	sitting	NaN	NaN
19	33H-AM-1019-02	wrestling with mother	NaN	NaN
25	2B-PM-1013-01	running (with nut)	NaN	NaN
29	6I-PM-1013-06	playing with #5	NaN	NaN
34	5F-PM-1007-03	hiding nut	NaN	NaN
37	20B-PM-1013-10	drank from a pond of rain water	NaN	NaN
43	28C-PM-1006-01	gathering acorns	NaN	NaN
50	41D-AM-1014-01	digging	NaN	NaN
53	38F-PM-1013-05	being chased,was pushed by other squirrel	NaN	NaN

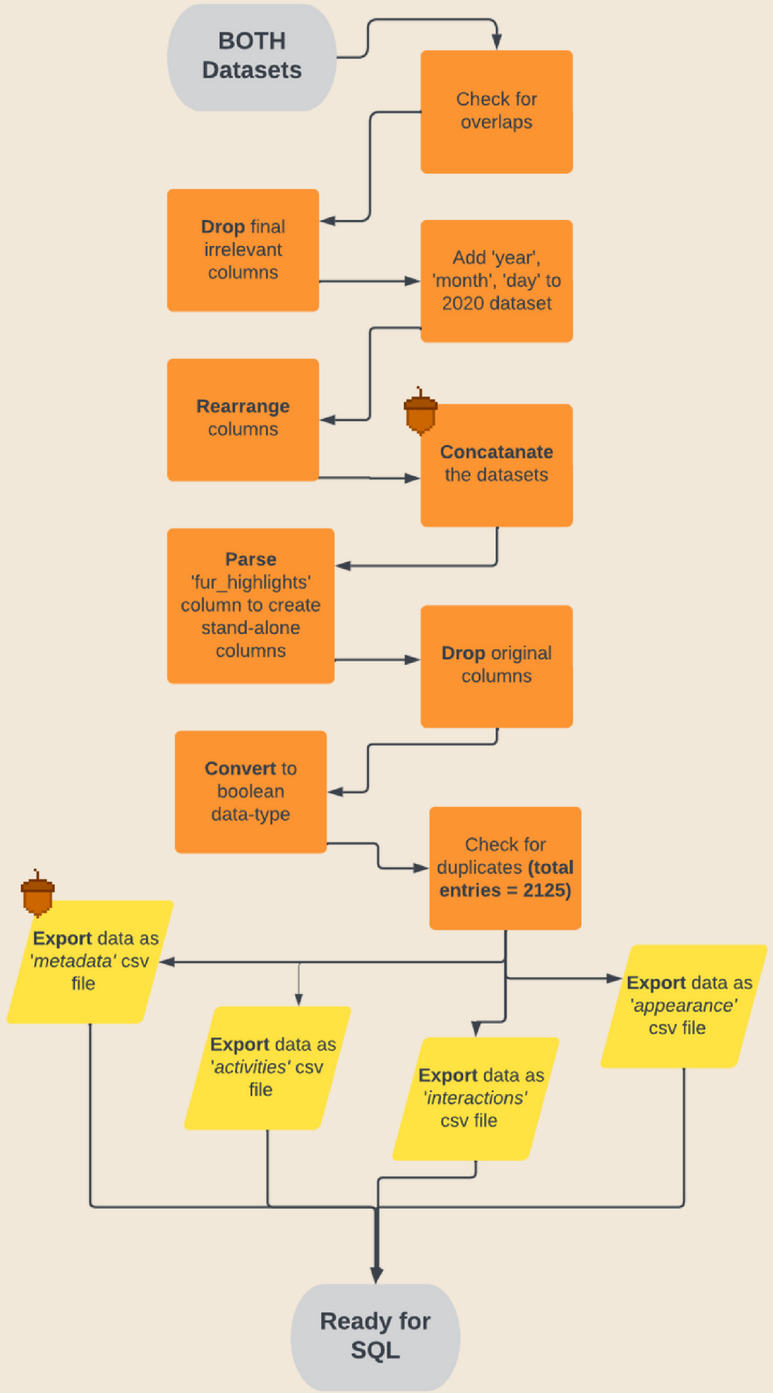
[94]:

```
# CONVERT THIS TO A DICTIONARY LATER, for simplicity!
# Combine activities
patterns = [
    'sit', 'watch', 'staring', 'posing', 'hanging out', 'chillin', 'dig', 'bury', 'chas', 'climb', 'eat', 'hop', 'foraging', 'running', 'jump']
activity_col = [
    'sitting', 'watching', 'sitting', 'sitting', 'sitting', 'sitting', 'digging', 'digging', 'chasing', 'climbing', 'eating', 'running', 'foraging', 'running', 'climbing']

all_indices = []
for idx_pattern, pattern in enumerate(patterns):
    idx_sit = list(other_df[other_df['other_activities'].str.contains(pattern)].index)
    # all_indices.append(idx_sit)

    for idx in idx_sit:
        all_indices.append(idx)
        other_df.loc[idx, activity_col[idx_pattern]] = True

# false_rows = set(other_df.index).difference(set(all_indices))
otherdf_cols = other_df.columns
for idx, row in other_df.iterrows():
    for col in otherdf_cols:
        if (row[col] == True):
            continue
        # elif (row[col] == "NaN"):
        elif isinstance(row[col], float):
            other_df.loc[idx, col] = False

other_df.head(15)
```

[94]:

	squirrel_id	other_activities	sitting	digging	watching	chasing	climbing	eating	running	foraging
2	11D-AM-1010-03	grooming	False	False	False	False	False	False	False	False
11	12I-AM-1013-01	sitting	True	False	False	False	False	False	False	False
19	33H-AM-1019-02	wrestling with mother	False	False	False	False	False	False	False	False
25	2B-PM-1013-01	running (with nut)	False	False	False	False	False	False	True	False
29	6I-PM-1013-06	playing with #5	False	False	False	False	False	False	False	False
34	5F-PM-1007-03	hiding nut	False	False	False	False	False	False	False	False
37	20B-PM-1013-10	drank from a pond of rain water	False	False	False	False	False	False	False	False
43	28C-PM-1006-01	gathering acorns	False	False	False	False	False	False	False	False


Please do go on


BOTH Datasets
Check for overlaps
Drop final irrelevant columns
Add 'year', 'month', 'day' to 2020 dataset
Rearrange columns
Concatanate the datasets
Parse 'fur_highlights' column to create stand-alone columns
Drop original columns
Convert to boolean data-type
Check for duplicates (total entries = 2125)
Export data as 'metadata' csv file
Export data as 'activities' csv file
Export data as 'interactions' csv file
Export data as 'appearance' csv file
Ready for SQL

Merge the datasets by concatanating them (one on top of the other)

squirrel_cleaning.ipynb

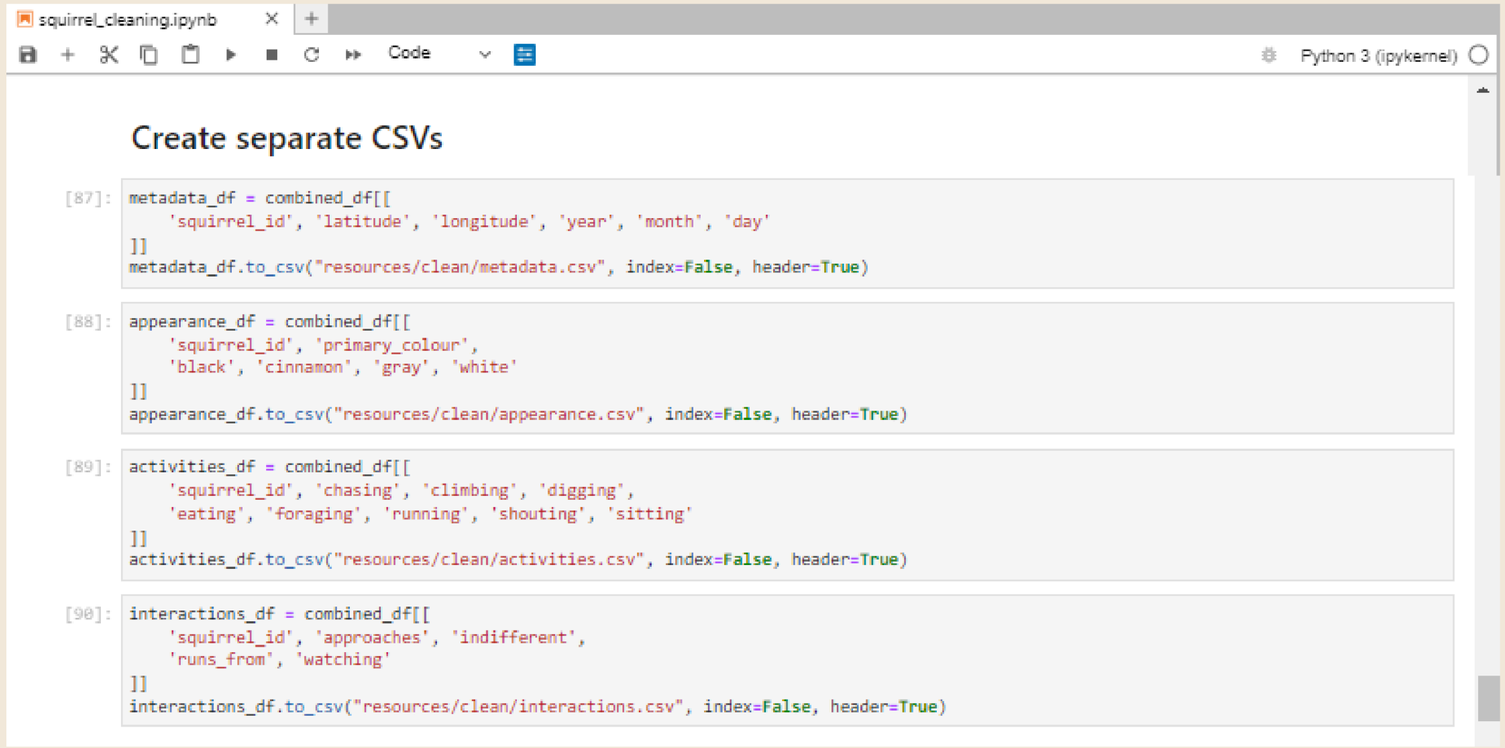
Code

[70]:
```
# Append the datasets
combined_df = pd.concat([s2020_clean, s2018_clean], ignore_index=True)
combined_df.head()
```

[70]:

	squirrel_id	longitude	latitude	year	month	day	primary_colour	fur_highlights	chasing	climbing	digging	eating	foraging	running	shouting	sitting	approaches	indifferent	runs_from	watching
0	A-01-01	-73.933936	40.859410	2020	3	1	Gray	White	False	False	False	False	True	False	False	False	False	True	False	False
1	A-01-02	-73.933937	40.859436	2020	3	1	Gray	White	False	False	False	False	True	False	False	False	False	True	False	False
2	A-01-03	-73.933894	40.859416	2020	3	1	Gray	White	False	False	True	True	False	False	False	False	False	True	False	False
3	A-01-04	-73.933895	40.859418	2020	3	1	Gray	White	False	False	False	False	False	True	False	False	False	True	False	False
4	A-01-05	-73.933590	40.859493	2020	3	1	Gray	Cinnamon	False	False	False	True	False	True	False	False	False	True	False	False

Export the cleaned dataset as four individual csv files ready for SQL



Final Visualisation

New York City Squirrels Dashboard

Project Overview

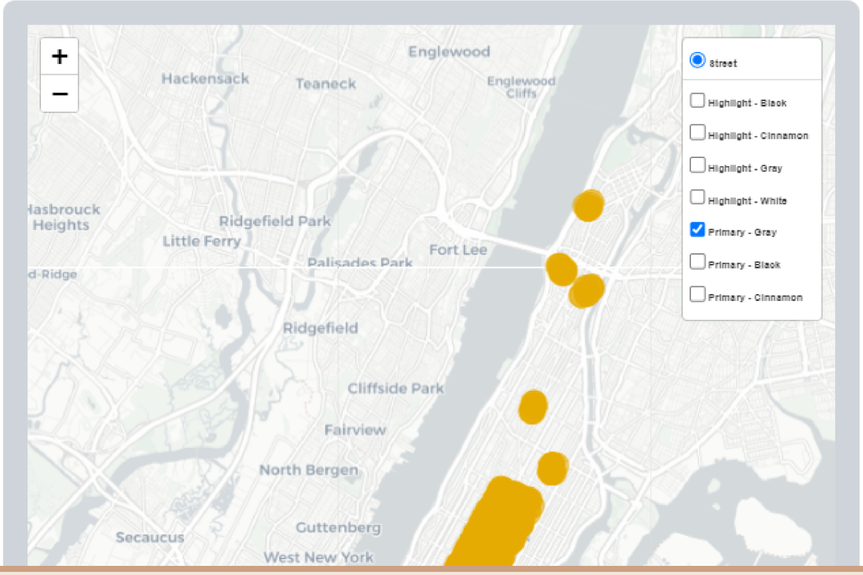
Datasets

Metadata

Appearance

Activities

Interactions



Spring Dataset | Autumn Dataset | Both Datasets

Activities | Appearance | Interactions

Key	Value
Squirrel ID	-
Date	-
Coordinates	-
Primary Colour	-
Highlights	-

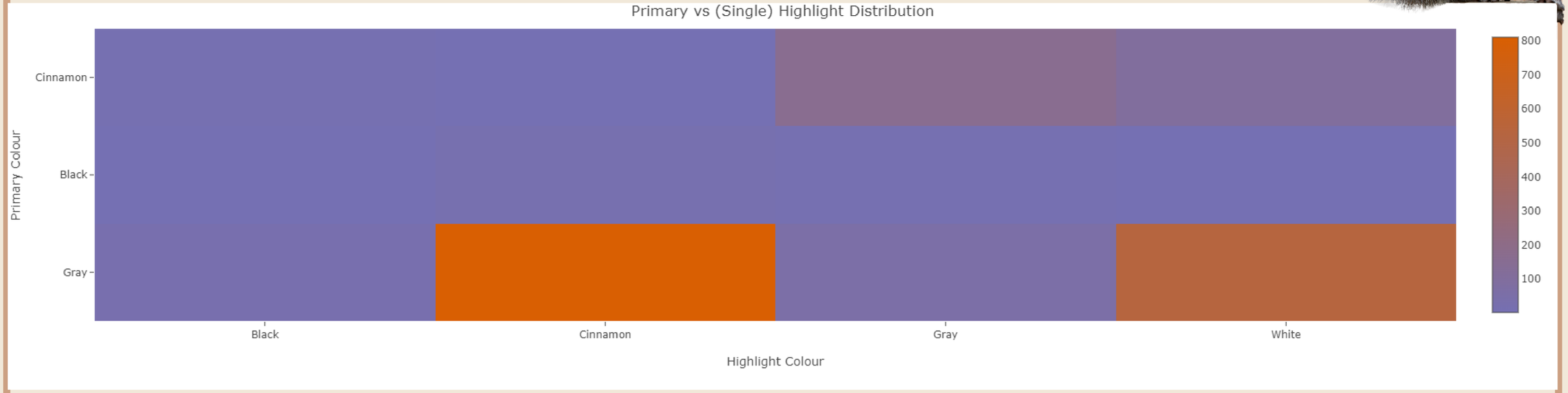
I had survive the hot summer the days of playing in the lake and seeking shade were behind me what a busy time it was.

I got to venture out of my nest and play with Gerald and Bobby, easy to tell apart by highlights cinnamon or white.

Weird cinnamons squirrels approached us the shady ones.
No wonder you can always see them hanging around the bench with their fancy caps on.

On the 5th day of the great sun Gerald and I thought we saw a black squirrel, we chased it and lost it up a tree Bobby says that we're lying and there's no such thing...

Squirrel Fur Colour Distribution



It is started to cool down we all need to be warmer this winter.

I've spent days digging a new food den behind the second big tree,
I don't think the others know it's there yet.

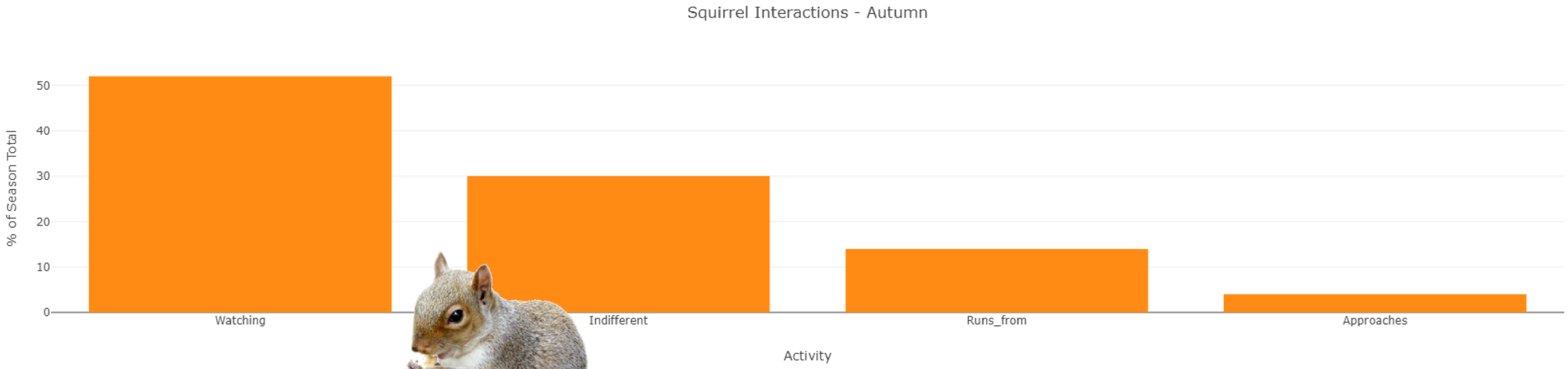
Quick! I have to make a run and get those big two acorns! Ah... success.

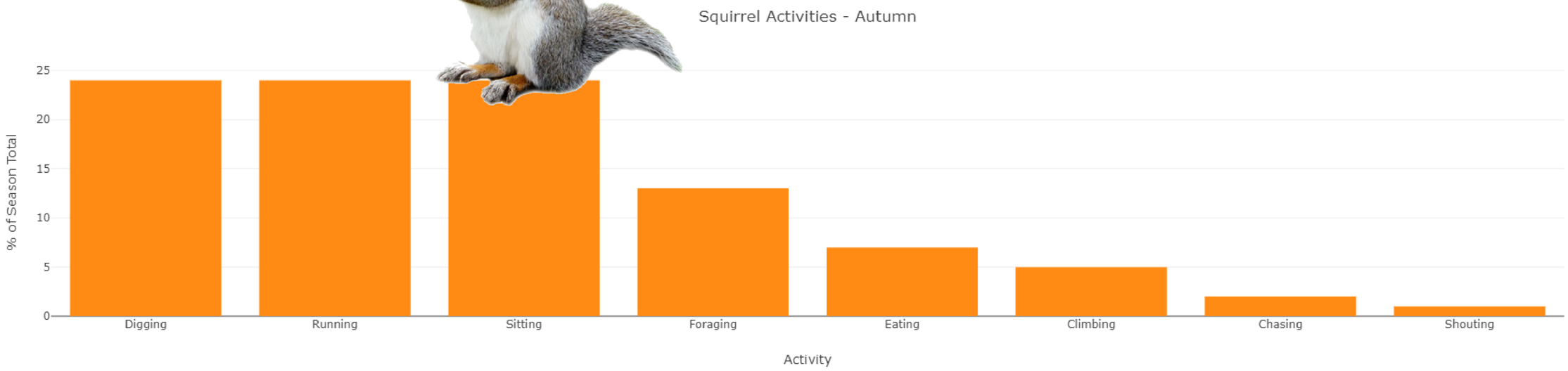
I've managed to fill four nut vaults and yet I still have to watch out for the big two legged ones... they're worse than the cinnamon squirrels.

Day three of digging and observing. I have had to run from the two legged ones after one of them approached me. I think it wanted my food or fur... unsure...

Gerald, Bobby and I will spend most of our time preparing for the winter approaching...

Life in Autumn





It is finally over.

The flowers are sprouting and new foliage begins to bud, spring is here!

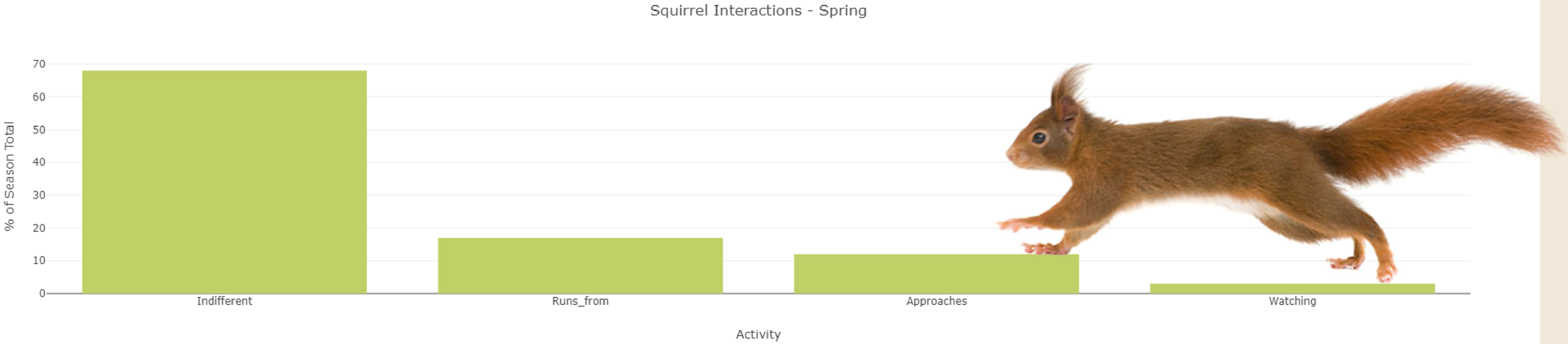
I am starting to find new sources of food, berries and flowers are coming back. Mhmm, tasty.

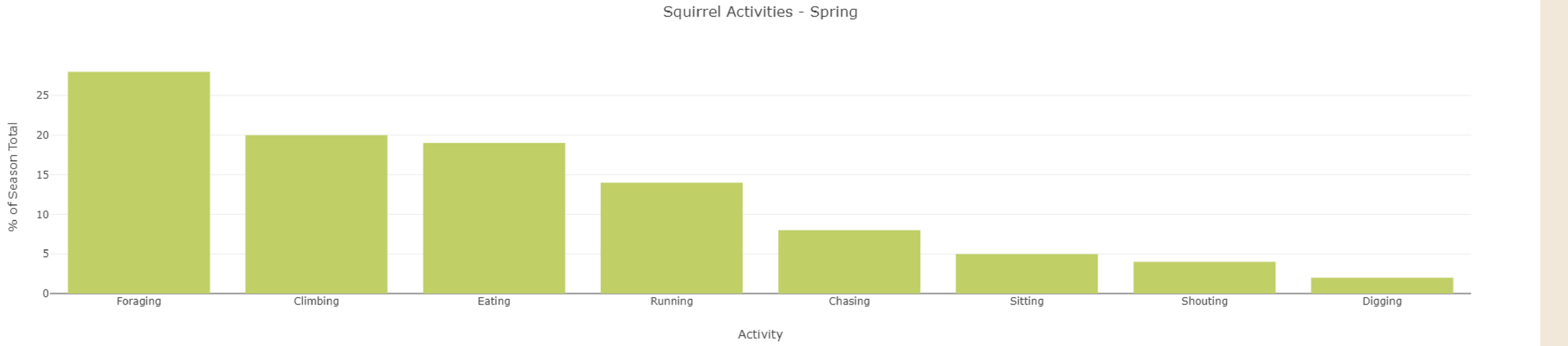
I have started to climb the second big tree and can see some berries. I will get to them.

I can see the 2 legged ones watching me, I don't care anymore, the mission is to climb this tree and get those berries...

Gerald managed to persuade the 2 legged one for a corn chip and has climbed another tree. Bobby and I need to try and get that delicious looking snack!

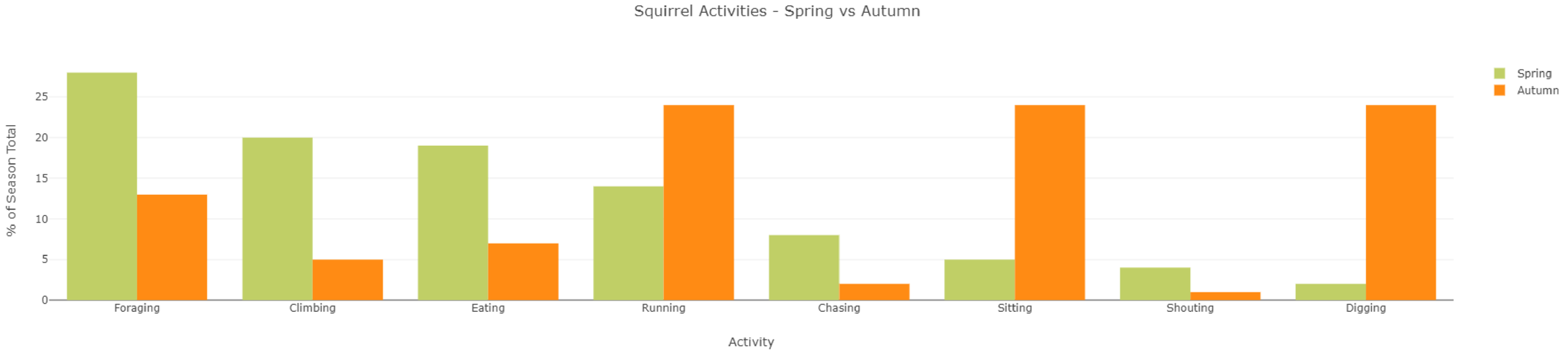
Life in Spring

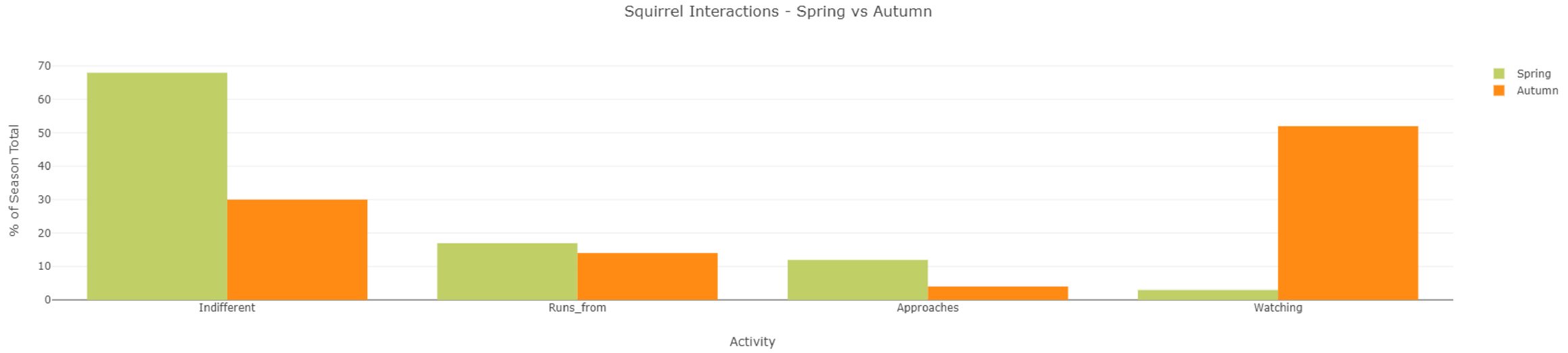




Question 1

How does the squirrel frequency and range of activities and behaviours compare between autumn and spring?


Squirrel Activities - Spring vs Autumn
% of Season Total
0
5
10
15
20
25
Foraging
Climbing
Eating
Running
Chasing
Sitting
Shouting
Digging
Activity
Spring
Autumn


Squirrel Interactions - Spring vs Autumn
% of Season Total
0
10
20
30
40
50
60
70
Indifferent
Runs_from
Approaches
Watching
Activity
Spring
Autumn

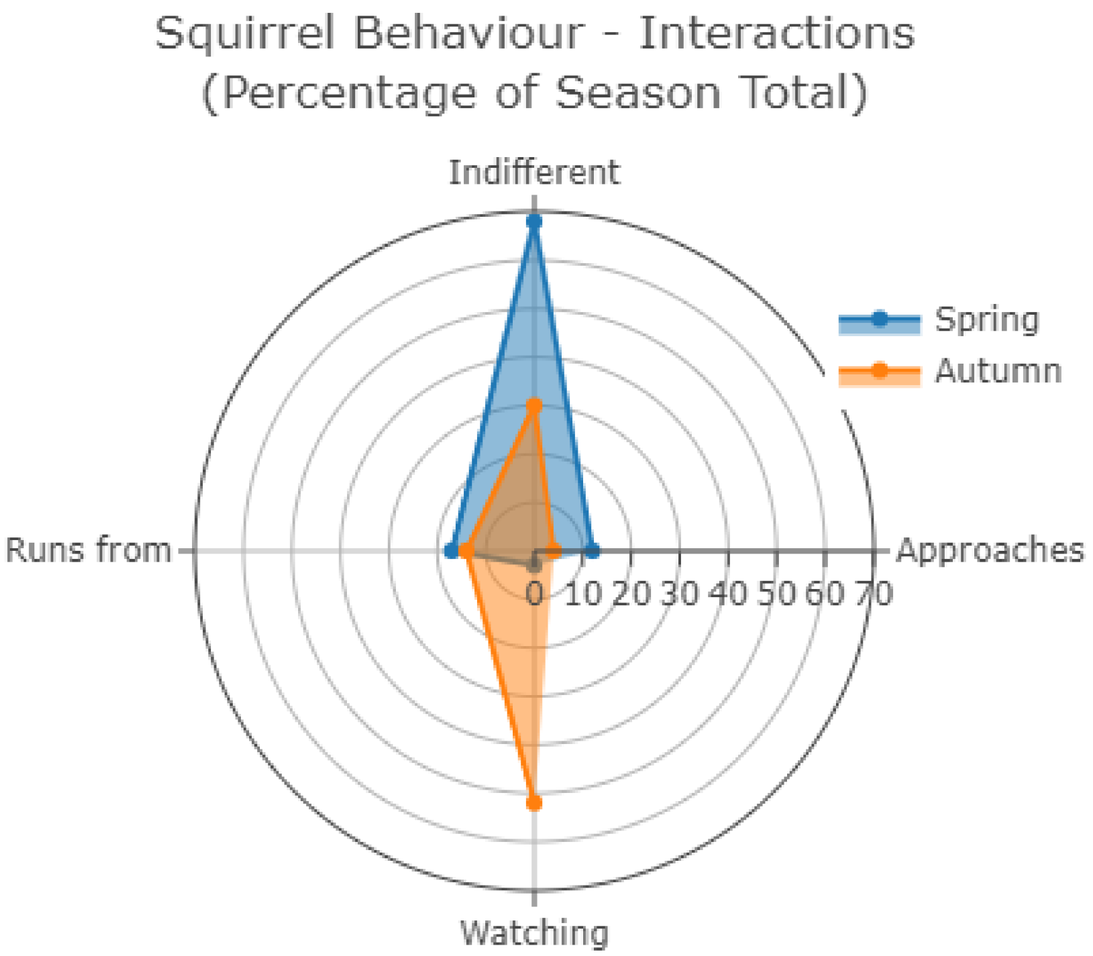
Question 2

How does the squirrel population diversity (colours, age, location) compare between autumn and spring?

No slide for this question - the conclusions are contained within the interactive dashboard.

Question 3

How do squirrels, in a park setting, interact with humans and other wildlife?


Squirrel Behaviour - Interactions
(Percentage of Season Total)
Indifferent
Runs from
Approaches
Watching
0 10 20 30 40 50 60 70
Spring
Autumn


THANK YOU