Squirrels in NY: Project Proposal

Team BCHZ - Nick Bayer, Richard Cui, Laura Han, Anna Zhang

Introduction

As a result of the continuous human development, animals are inevitably interacting with humans more often. However, this form of interaction has mostly shown to be a disturbance to animals [1]. Animals see humans as a threat, so it is no surprise that they would treat the presence of humans the same way they would when they face other predators. Nevertheless, recent studies show that the squirrels actually act differently, as characterized by a phenomenon called synurbization, or the process of becoming urbanized [2].

In an effort to investigate these two competing theories, and to better understand the dynamic between squirrels and humans, we carry out this research project to explore what factors affect whether a squirrel is indifferent to human presence. From there, we would like to deduce whether the squirrels' attitude to humans are caused by human presence or other factors such as their species.

We hypothesize that the age category, location, distance above ground when spotted, number of activities that the squirrel was doing, sound that the squirrel makes, and whether squirrel is disturbed by human activities (as measured by features like approaching or running away from humans and tail signs) could have a relationship with the attitude of the squirrel (whether indifferent or not).

Data description

We are sourcing our data set from the TidyTuesday project on GitHub. Their data originally came from The 2018 Squirrel Census, a project based on the sightings of squirrels in Central Park, New York City.

In October of 2018, the Squirrel Census Team and a group of over 300 volunteers collected the data based on squirrel sightings around Central Park.

The dataset has 3023 observations and 31 variables, and it gives a wide range of observations and characteristics. It first gives us the location, in both longitude and latitude, the hectare of the park the squirrel was located in, the date, and whether it was found in the AM or PM. It

also assigns each squirrel a unique ID. It has information on whether the squirrel is an adult or a juvenile, its primary and highlight fur colors, and has a number for the sequence of sightings in one session. It also contains data on their exact location, their distance from the ground, and the objects they were found on. There is data for the activities the squirrel was found doing, ranging from running to foraging, with a separate column for any activity that was not chosen to be a column. It gives data on the sounds the squirrel made and tail movements, if any. Finally, it has 4 columns for the response, being either that the squirrel approached, was indifferent, ran away, or any other action.

Initial exploratory data analysis

```
squirrels <- read_csv("data/squirrel_data.csv")
```

We create a numeric variable named Activity_Score that encapsulates the number of activities a squirrel is engaged in during the span of observation. The distribution of the Activity_Score variable is unimodal, slightly right skewed, and has a median of 1. This means that most squirrels in the data set were only engaged in 1 activity during the span of observation.

```
activities <- squirrels |>
  select(c(Running, Chasing, Climbing, Eating, Foraging)) |>
  mutate(Running = ifelse(Running == TRUE, 1, 0),
         Chasing = ifelse(Chasing == TRUE, 1, 0),
         Climbing = ifelse(Climbing == TRUE, 1, 0),
         Eating = ifelse(Eating == TRUE, 1, 0),
         Foraging = ifelse(Foraging == TRUE, 1, 0))
squirrels$Activity_Score <- rowSums(activities)
```

```
dist_activity <- squirrels |>
  ggplot(aes(x = Activity_Score)) +
  geom_bar(fill = "brown") +
  labs(title = "Spread of Acitivity Scores",
         x = "Activity Score",
         y = "Frequency")
summary(squirrels$Activity_Score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   1.278   2.000   5.000
```

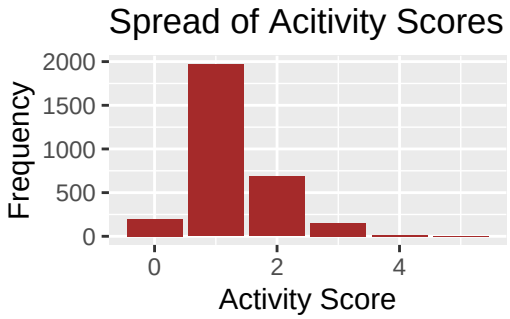
```r
dist_response <- squirrels |>
  ggplot(aes(x = Indifferent)) +
  geom_bar(fill = "brown") +
  labs(title = "Distribution of Indifference",
         x = "Activity Score",
         y = "Frequency")
```

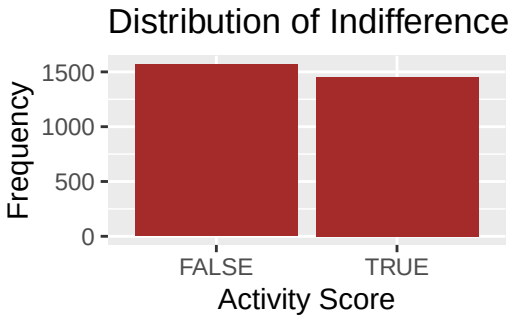
```r
dist_approach <- squirrels |>
  ggplot(aes(x = Approaches)) +
  geom_bar(fill = "brown") +
  labs(title = "Distribution of Whether or not Squirrels Approach Humans",
         x = "Whether Squirrel Approaches",
         y = "Frequency")
```

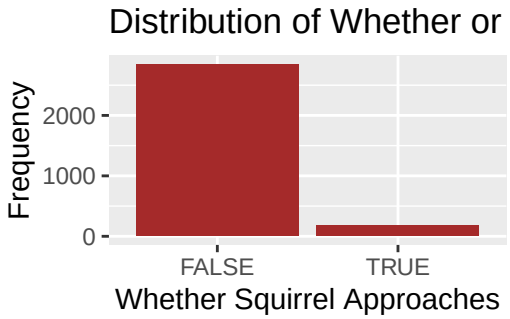
```r
indiff_activity_graph <- squirrels |>
  ggplot(aes(x = Activity_Score)) +
  geom_bar(fill = "brown") +
  facet_wrap(~ Indifferent) +
  labs(title = "Indifference vs. Activity Score",
         x = "Activity Score",
         y = "Frequency")
```

```r
approach_activity_graph <- squirrels |>
  ggplot(aes(x = Activity_Score)) +
  geom_bar(fill = "brown") +
  facet_wrap(~ Approaches) +
  labs(title = "Approach vs. Activity Score",
         x = "Activity Score",
         y = "Frequency")
```

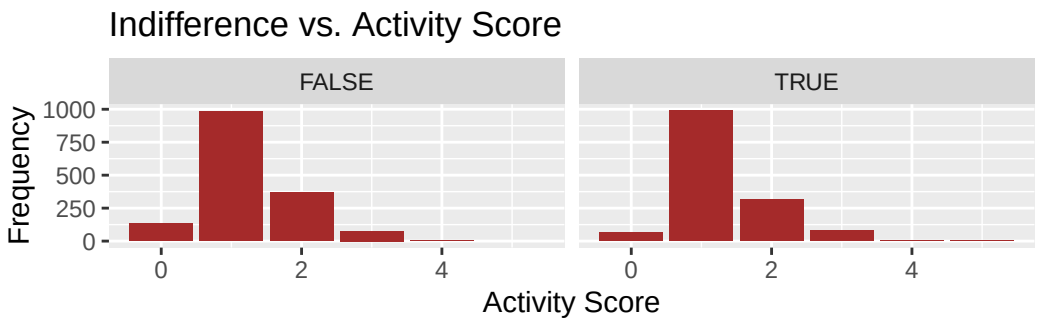
```r
grid.arrange(dist_activity, dist_response, dist_approach, nrow = 2)
```

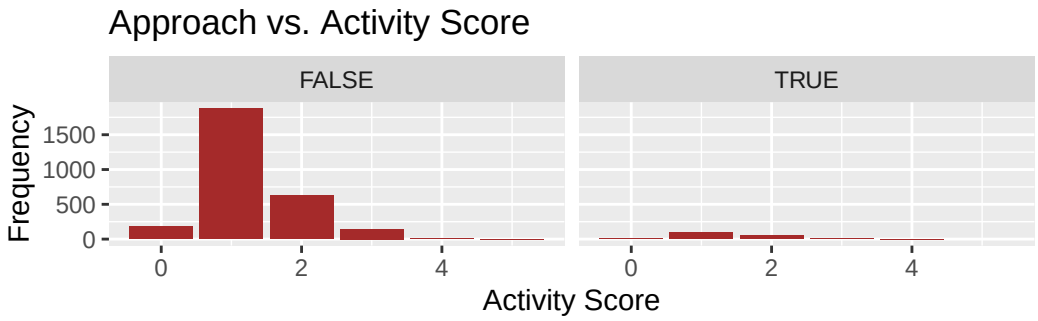







```
grid.arrange(indiff_activity_graph, approach_activity_graph, nrow = 2)
```





There seems to be roughly equal number of squirrels that were and weren't indifferent to the researchers.

There were far fewer squirrels that approached the researchers than squirrels that didn't. This could show that they are not as synurbanized as we thought.

The distribution of Activity_Score for squirrels that were indifferent vs. not indifferent is roughly the same shape. Therefore, activity score and indifference do not seem to be correlated.

Since there were so few squirrels that approached the researchers, it is hard to tell whether the graphs are very different in shape. The difference may be something that we explore further.

An interaction that we would like to examine further is whether the presence of tail flags and/or twitches influences the likelihood that a squirrel is indifferent to the researchers. Another interaction to explore is how the presence of tail flags plus the number of activities would affect the squirrel's indifference towards humans.

Analysis approach

The response variable is `Indifferent`, which is a categorical variable that indicates whether or not the squirrel is indifferent to humans or not. Potential predictors include `Activity_Score` (a quantitative variable that records the number of activities the squirrel is observed doing), `Age` (a categorical variable that indicates whether the squirrel is adult or juvenile), `Fur_Color` (categorical variable), `Location` (categorical variable), `Above_Ground_Measurement` (quantitative variable), sounds that the squirrels are making (categorical variables including `Kuks`, `Quaas`, and `Moans`), `Tail_flags` and `Tail_twitches`, which are also categorical.

To explore the relationship between whether or not the squirrel is indifferent and the predictor variables, such as age category, location, distance above ground when spotted, number of activities the squirrel is observed doing, sound that the squirrel makes, and their tail signs, we plan to use multiple linear regression. We will compare multiple linear regression models using R^2_{adj}, AIC, and BIC to evaluate what predictor variables and what interactions between the variables should be included in the model to best predict the attitude of the squirrel towards humans. We will also perform 10-fold cross validation to assess the performance of the model and estimate its generalization ability.

Data dictionary

The data dictionary can be found here.