STA2201 Research Project

Investigating the Characteristics of Courageous Squirrels

Timothy Jordan Regis

2023-04-20

Intro

Research Question and Motivation

Common tree squirrels have often been considered to be quite timid creatures, with a strong tendency to flee from approaching humans or animals, which makes good sense when we acknowledge their size. Despite this, however, there have been many cases in which these squirrels do not show this same level of fear, and even some in which the squirrels show curiosity or other domesticated qualities. This brings to light a question which asks why some squirrels can be less fearful, and if this is a result of their past experiences, or potentially related to genetic differences across the species. As such, the goal of this paper is to develop a better understanding of the interesting habits and characteristics of squirrels, using statistical techniques. Specifically, I aim to understand the factors which make squirrels more or less likely to be afraid of humans.

My fascination with these creatures developed in high-school, when a squirrel sitting in my driveway was unshaken by our car pulling up, and showed no signs of caution as we walked past it and went inside. I went back out after noticing this and found the squirrel allowed me to approach it and was even more curious than I was when I reached out a finger, as it investigated to determine if it were food. I was surprised to see a squirrel so confident, as many in the past were often quick to avoid any approaching humans or animals, and I was never able to understand how that could be. Following this encounter, the squirrel returned a few times, which eventually grew into the squirrel breaking into our house through a screen window, an attempt that was fortunately thwarted by our resident feline. But this event stuck with me for quite some time as I realized that some of these animals could have very different personalities, despite their size. Thus, upon discovering that there were people that had collected data I could use to answer this question I became almost instantly attached.

Previous Work

There is a surprising amount of past work in this field that paints somewhat of an already intuitive picture of the behaviour of squirrels. In the paper "Behavioral Responses of Eastern Gray Squirrels in Suburban Habitats Differing in Human Activity Levels" (Cooper et. al. 2008), the authors observe the responses of squirrels based on either a lone human, or a human with a leashed dog, and naturally found that the squirrels were more alert when in the presence of a dog, and were more likely to flee with greater levels of human activity surrounding them. The article "Responses of urban gray squirrels (Scurius carolinensis) to humans and conspecifics in an area of Boston Common" (Rademacher, 2018) took this paper another step further by investigating a smaller group of squirrels over a longer period, finding that squirrels were more likely to flee from louder and more aggressive humans, but can often be tempted into approaching a human when they are more relaxed or offer some food. In the paper "Shyness and boldness in squirrels: risk-taking while foraging depends on habitat type" (Nowak et al. 2018), the authors again found that squirrels are more 'bold' when in urban environments than in forests, through the use of multiple experiments. While these results are quite valuable, they all focus on a much smaller group of squirrels, and tend to only focus on one main factor, which is primarily environmental. As such, our paper aims to answer this question on a much larger scale, as well as use more detailed characteristics of squirrels to attempt to explain their behaviour further.

Data

For this paper, I have used data courtesy of 'The Squirrel Census', a volunteer-supported census, founded in Atlanta, tracking the behaviours of Sciurus Carolinensis, the common Eastern Grey Squirrel. The organization has a deep affection for squirrels and focuses heavily on explaining the stories behind each encounter while also tracking the data in a standardized format. Specifically, we will be using data from 2 of these censuses. The first of these comes from a census in October of 2018, where 323 volunteer Squirrel Sighters visited Central Park in New York City, where they surveyed the area for squirrels over the course of a couple days, observing over 2,000 squirrels. This dataset is provided through the NY Data Portal, and was fortunately hosted by **rfordatascience** on GitHub and featured on tidytuesday, potentially allowing for an easier method for sharing this project. The second set is from their March 2020 census, where this time 72 researchers went to 24 different parks in New York City and again surveyed the area for squirrels. This dataset has been provided directly through the Squirrel Census website, which features links to Dropboxes containing their data from this census. This second set was our original data, however, due to a lack of significant relationships found in preliminary regressions, we decided to incorporate a greater span of data to hopefully achieve more satisfactory results. In both of these censuses, the sighters marked down qualities of the squirrels, like their fur color and highlights, the location of the squirrel when spotted, its current activity, and most importantly, their interactions with humans and whether they were indifferent to their approach or if they would run from them.

We do, however, have some concerns with the data itself that we will briefly discuss here. Firstly, there is some natural suspicion that the volunteer sighters may have recorded data on the same squirrel more than once without knowing it. Say, if a squirrel had run on its first encounter, and near the end of the survey period that squirrel is seen again and recorded as a new observation. Unfortunately, there is little we can do to alleviate this issue, meaning that we must move on with it, but keep this suspicion in mind when we are trying to understand our model. Second, if squirrels truly are very timid creatures, then spotting them could actually be quite difficult as they may be quick to run away before humans even get close. If this were true, this would suggest to us that our data is really only covering squirrels that 'want' to be seen, in that they are already less afraid than most squirrels, which could heavily bias our model. Again, however, there's really nothing we can do to fix this issue ourselves and thus keep it in mind for the future. Lastly, the general concern we have with the data is the variance behind each observer, as humans can be surprisingly different in the ways they interpret the things they see, which we worry may show up in the data. For example, how a person records a squirrel's interaction and at what point they do this can make a big difference, and if we aren't careful with these differences then we may have a completely unreliable model. We can attempt to alleviate this by grouping the data in various ways, such as by some type of geographical aspect. The best way to handle this would be if each observation had an indicator for the ID of the spotter themselves, but this was unfortunately not provided in the data, thus again we move forward with our data as such, keeping in mind all of these worries as we carry on.

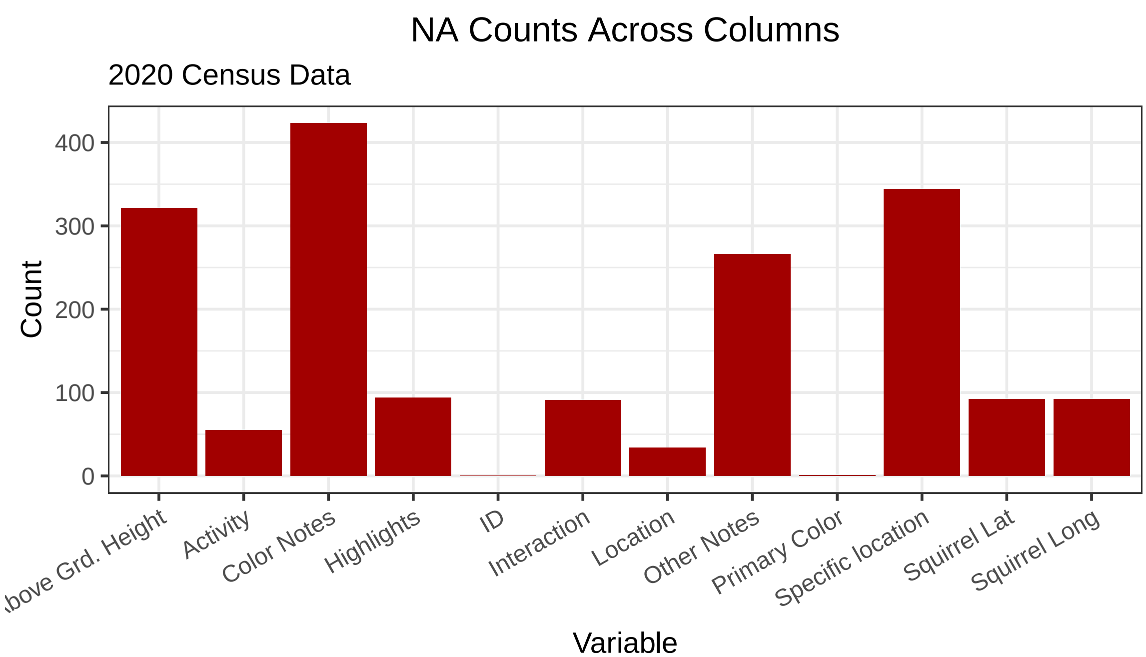
EDA

Issues & Cleaning

To begin with our exploratory data analysis, we must first do the job of data cleaning, and, as our data has been manually filled out by a number of volunteers, there is a lot that needs to be done before we can begin our analysis.

Table 1: NAs from 2018 Census

	NAs
age	121
primary_fur_color	55
highlight_fur_color	1086
color_notes	2841
location	64
above_ground_sighter_measurement	114
specific_location	2547
other_activities	2586
other_interactions	2783
zip_codes	3014



To begin with our EDA, we first notice that there are many NA values filling both datasets. Firstly, in our 2020 data, things like 'Specific Location', and 'Color Notes' are almost entirely NA. However, these columns are tracking notes made by the sighters on each spotting, and are unhelpful for the purposes of our analysis, so they can be safely ignored. Despite this, in the column tracking the squirrel's interaction with humans we have 91 missing values, 34 tracking the location of the squirrel, and 55 activities. Unfortunately, however, there is little we can do here aside from removing them, as no information can be used to impute the data. For the Highlights in Fur Color column, we get many NAs as well, but here we believe these NAs can be imputed as squirrels with no highlights in their coat, and only identifiable by their primary color as this is quite a common occurrence.

Next, in **Table 1** we can see that again there are many NA values coming from the 2018 census as well. Again, however, we notice that these show up primarily in columns such as 'color_notes', or titles beginning with 'other', which are only used for additional commentary and will not be helpful for this analysis. We do notice some missing cases in important variables such as highlights and fur color, but these are more likely due to non-reporting, and we make the same decisions as in the previous dataset to deal with them.

After cleaning these NAs, we next observe some of the values recorded in our predictors, noticing yet another problem with the use of volunteer recorders.

Table 2: Activity Recordings

Foraging	Eating, Foraging	Climbing (down)	Climbing, Balancing on fencing
Eating, Digging something	Climbing, Eating	Running, Eating (peanuts)	Chillin', Rubbing butt on ground
Running	Running, up tree	Climbing, Watching #2	Sticking out of a tree
Running, Eating	Sitting on branch	Running, Chasing, Eating	Hangin' with #13 & #14
Climbing	Sitting at attention	Climbing, Eating, Foraging	Climbing (tree)
Eating, Digging	Sitting, shouting	Running, Chasing	Snacking in a tree
Foraging, Nesting/gathering leaves	Defending tree, shouting	Running, Climbing, Eating	Prancing about
Chasing	Cleaning	Ear scratching	Climbing fence
Eating	Resting in tree	Frolicking	battery
Sleeping	Running, Climbing, Foraging	Running, Climbing, Scratching	Climbing, Foraging, Self-cleaning
Sitting	Running, Eating, Foraging	Running, Climbing	Climbing, Sitting in short tree
Chasing, Climbing	Digging	Posing	Foraging, Jumping
Chasing, Climbing, Eating	Running, Digging	Guarding	Chilling
Running, Chasing, Climbing	Grooming	Running, Eating (or pretending to eat)	Hanging
Climbing, Foraging	Nesting	Eating, Burying	Eating (bread crumbs), Foraging
Vocalization at us	Climbing (down tree)	Very carefully watching a cat	Hanging out
Running, Foraging	Sitting (in tree hole)	Chattering	Sleeping (Dead?)
Jumped to building	Lounging	Eating (nuts), Foraging	Eating (tortilla/chip)

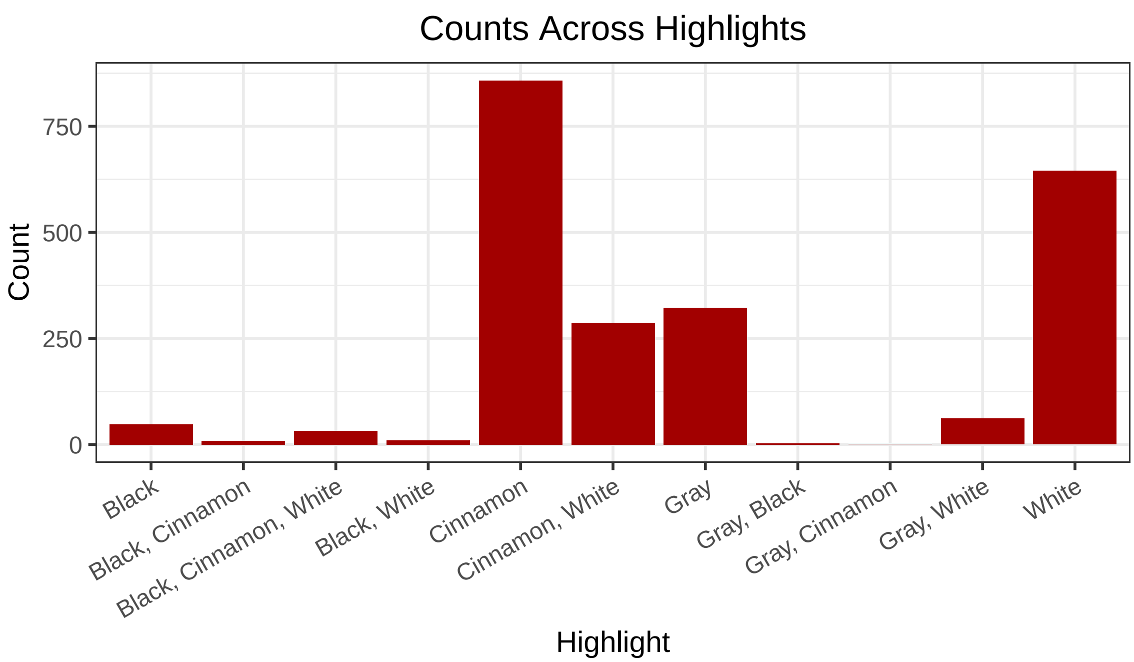
Looking at the results from the activities variable in **Table 2**, we can see that many observations do not follow a clean format of one-word terms to describe an activity. Instead, many squirrel sighters have either used multiple terms to classify the activity, potentially pointing to them observing the squirrels over a longer period, or they'll use a much looser nomenclature to describe their activities, such as "Hangin' with #13 & #14". In the documentation, we can find that this is because the column also includes any additional notes made by the spotter. Thus, to solve this we had two main options; we could cut off all phrases after the first key term is used, saving more or all of our data, or we could remove all non-singular terms altogether, which would reduce our total available data.

To save more of our data, however, we decided to stick with the first option. We believe that the secondary terms could be done after their first interaction, leaving open the possibility that a sighter sees the squirrel again after marking down its interaction on their first encounter. We do, however, see that many of these activities are recorded with only one observation, and we have chosen to remove these recordings as they are likely the result of a volunteer having a much different approach to filling in their sighting details than the rest of the group.

After this operation, we still found that many of the recorded activities were quite similar, such as "Chilling", "Hanging", "Lounging", etc. In turn, we decided to group these observations based on the main trait behind the activity, for the aforementioned case we chose "Relaxing" as the overarching theme to these activities, and we believe this will allow for a more concentrated look at squirrel behaviour and potentially remove some of the inherent variances due to the natural differences of the volunteers. Furthermore, this gives our model larger pools of data to pull from which should give us stronger and more accurate estimates of our parameters.

In the 2018 census data, however, this issue is fortunately far less prevalent, as the recording for activities is done by indicator columns for each of the main activities, meaning that we can simply pivot these columns into a single column to tell us the given activity of each squirrel.

After, cleaning the activities column, we found a similar issue present in the highlights column.



Here we can see that many of the squirrels have somewhat unique highlights, such as (black, cinnamon, and white), apart from the main groups of white, grey, and cinnamon. The issue here, however, is that we can't determine if these are due to the sighters having different interpretations of colouring, or if there are actually rarer color schemes for squirrels that exist in New York. As such, we have decided to leave the variable untouched aside from removing NAs as we believe it gives a more honest approach to the data.

A similar issue was also found in our location variable, and we performed the same operation of choosing the two base options; Above Ground and Ground Level.

Lastly, in our interaction variable, this problem came up again as expected at this point. The two main reactions we are interested in are Indifferent, and Runs From, thus, to clean this data up, we filtered to only include observations that used exactly these responses, so as to not disrupt our actual outcome variable. Furthermore, upon inspecting the count of each reaction, we found that the names different from the "Indifferent" and "Runs From" were almost all seen only once in the data, leading us to again believe that these are cases of squirrel sighters with highly unusual naming conventions, making us more confident in removing them from the group.

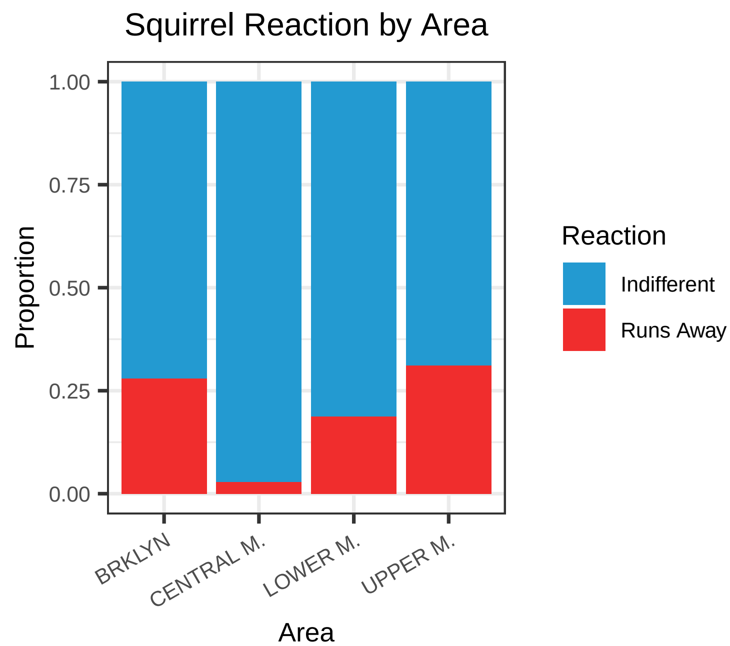
Census	Prop. of Runners
2018	0.3155
2020	0.1055

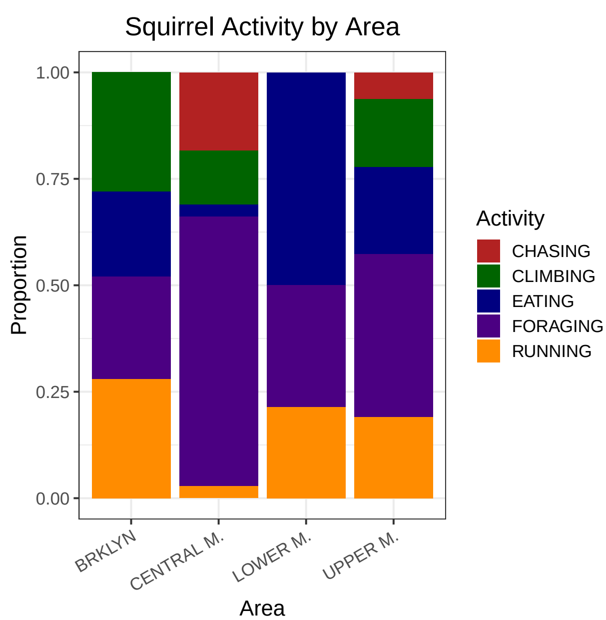
After finally cleaning the data, we arranged the columns of both datasets, matching up the naming and text conventions in each, and joined them together to create the final dataset we intend to make inference from.

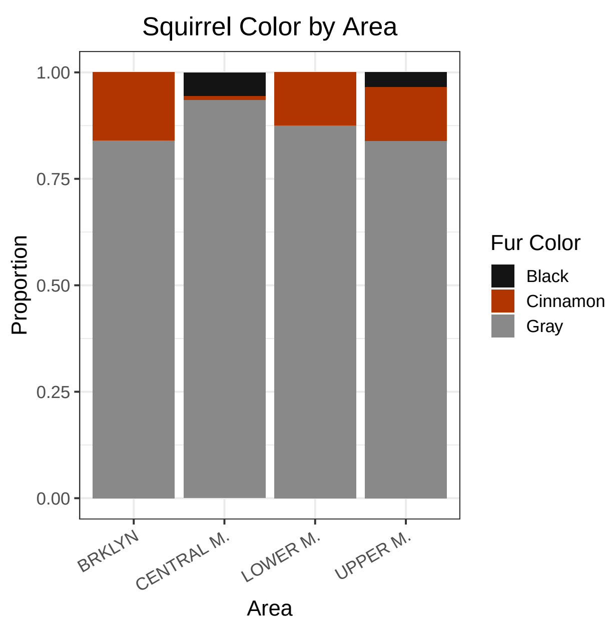
To ensure that combining the two datasets would not disrupt too much of our model, we first checked the distribution of the response variable in each setting, "interaction", to which we can see 30% of the population running in the 2018 census, and just 10% running in the 2020 census. While these are fairly different, we believe that combining them should not cause too significant of an issue, as both proportions are well below 50% signifying that they both agree with the same trend that the majority of the squirrels seen tend to remain indifferent to a human's approach. This may, however, make us interested in applying a hierarchical structure based on the census data of the observation, or including this factor as a covariate to our model.

Visualizations and Preliminary Results

Now that we have our data finalized, we can take a look at some of the interesting trends that exist.



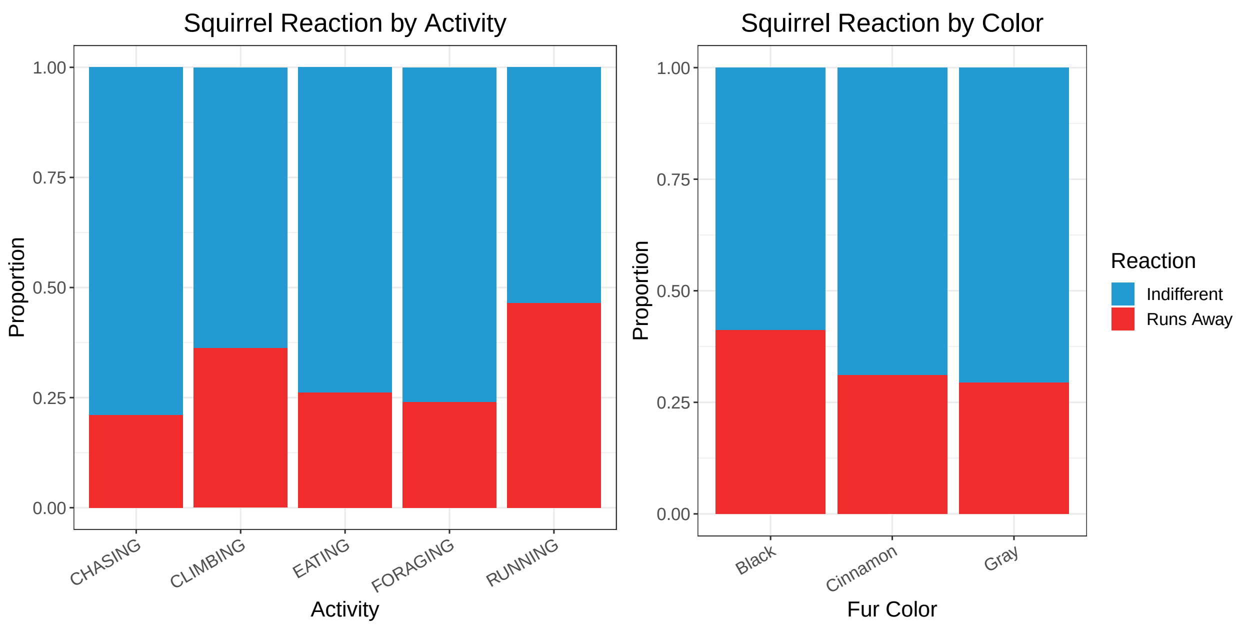




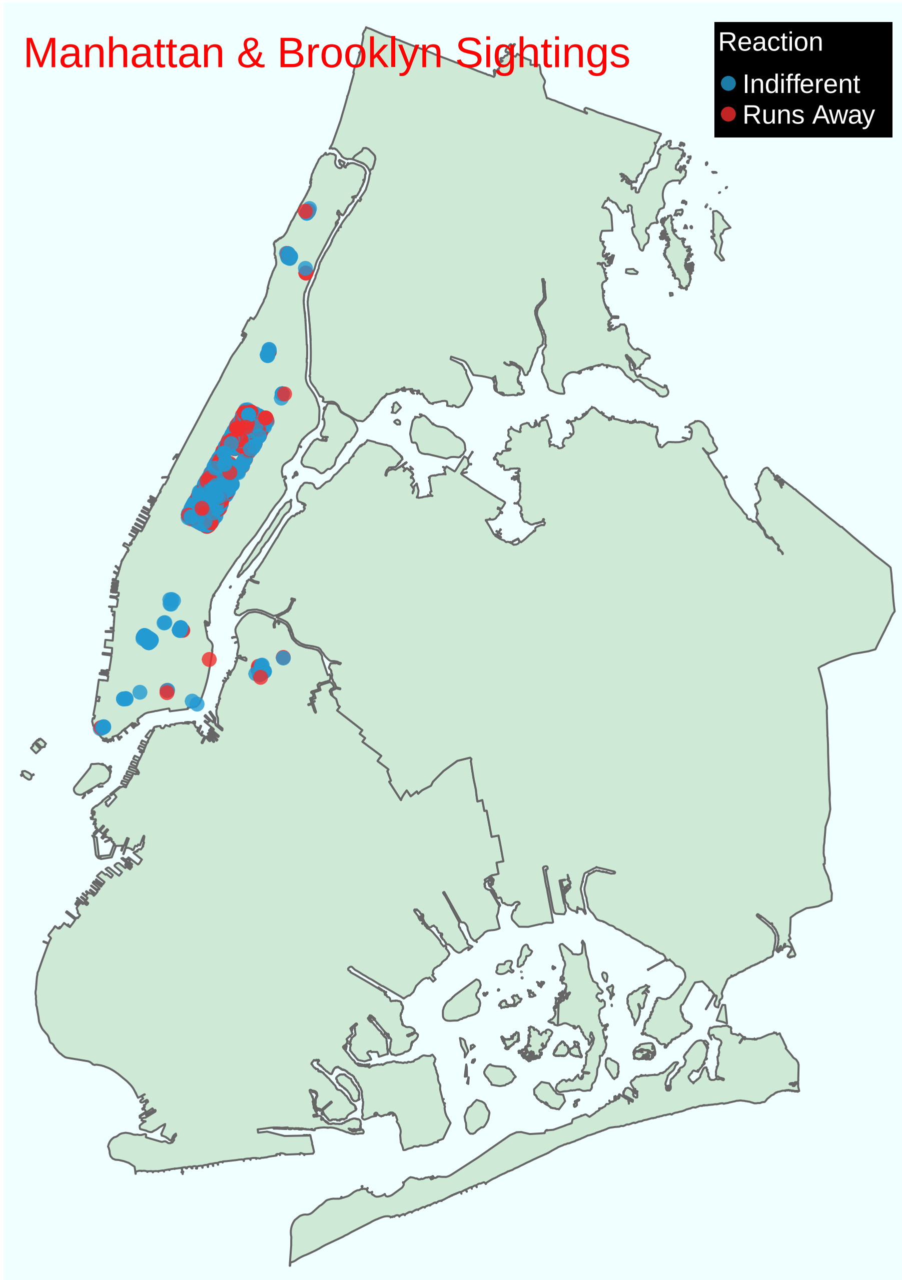
In our first plots, we examine the breakdown of some of our key variables by the area the squirrels were spotted in. First, we examine the breakdown of their interactions with humans. We can see here that a majority of these squirrels actually appear indifferent toward humans, while fewer are running off when approached. This is somewhat surprising given their commonly considered timid nature, however, this may point to the fact that many of the squirrels seen in the data actually 'want' to be seen, in that the squirrels more afraid of human interaction are not even out at this time, and instead frequent the parks when most humans have gone to sleep.

We can also see that Central Manhattan has a much lower proportion of squirrels running away than the other main areas. The main result of this is relatively unclear, however, when we compare this result with the breakdown of our other variables we can begin to see some interesting patterns arise. Firstly, when we look to the plot of activity by area, we can notice that Central Manhattan also sees the largest number of squirrels currently foraging in the data. This may point to a relationship that suggests that squirrels in the middle of foraging for food are more likely to disregard potential danger. Conversely, we see the other areas with the higher proportions of runners, as well as the higher proportions currently eating, leading us to believe that more leisurely tasks may be more likely to be forgoed in the event of an approaching human.

We can also see a breakdown of the colorings of squirrels in the data, noting the very high proportion of grey squirrels across all areas, which was to be expected given the species that's being investigated. While there are still some differences in the color representations seen, this may point us in the direction of incorporating information about the squirrel's highlight colors in their coats to predict their behaviours more accurately. Interestingly, we do not see any black squirrels in Brooklyn or Lower Manhattan. We can also see that the areas with greater numbers of cinnamon squirrels see higher proportions of runners, potentially pointing to a relationship here that we will investigate further.



When we plot reaction against activity, we can see a relatively even distribution of runners over most groups, except climbing and running, which may suggest that the more 'active' activities are traits of squirrels that are more skittish and likely to flee to safety. Interestingly, when we then break down squirrel reaction by our fur color variable, we notice that black squirrels actually have the highest proportion of runners, which is somewhat counterintuitive when we look to Central Manhattan with the lowest proportion of runners yet highest proportion of black squirrels. Thus, it appears there may be a more confusing relationship than expected here, which we hope to uncover as we progress.



Our last plot displays the geographical locations of these squirrels in Manhattan and Brooklyn, colored by their interaction with humans. For every spotting, the volunteers were also able to track the latitude and longitude of the squirrels, allowing us to visualize these patterns across New York. While we don't see any clear patterns here, within the 2020 census data (not the large rectangle that is Central Park) there is still a pretty strong tendency for the squirrels to congregate near areas lining the water, likely due to the increased resources available to them, leading us to want to use geographical controls when building a model. Furthermore, we can see slightly more 'runners' closer to the water, potentially suggesting that the more 'comfortable' and 'fearless' squirrels move further into the city as time goes on. When we look at the Central Park data, there is little to be attained from this due to the density of points, and the sheer volume of sightings in comparison to the 2020 study.

Model

Now that we've gotten a handle on our data, we can begin with the modelling process.

Firstly, the variables we want to study in this model include; primary fur color, highlight color, activity, and area. We were somewhat limited by the number of variables matching between each dataset, but we don't consider this to be a huge issue as we only lose out on one potentially important variable in the 2018 census data, squirrel age. However, we are fine with this as the estimation protocol for age is unclear to us, and we have some doubts about the likelihood that all *323* squirrel sighters had the same level of accuracy throughout the study.

The other variables we had access to but did not select for this study included squirrel location and park name, and we will briefly touch on the reasoning here. Firstly, for squirrel location, in preliminary runs, we found extremely poor significance across all experiments, and saw little to no relationship in our visualizations. For park name, we have chosen to classify geographical trends by their area instead, each park was investigated by a different group of sighters, and we believe that fixing our effects on these groups would impart too much of their individual beliefs into our model, whereas by grouping by a few parks per area, we hope to smooth out these differences when estimating. Furthermore, there are 25 total parks included, which would massively increase the number of covariates in the model, compared with the 4 areas given in the data which can reduce this effect as well.

To answer the question of a squirrel's predicted interaction, we created an indicator variable that tracks whether a squirrel was indifferent or if they would run away from humans and used this as the dependent variable in the model. Naturally, with a binary outcome, we will be using a logistic regression model to predict the odds of a squirrel running from a human.

As all of our covariates are categorical as well, we also have many options for bringing our model to a hierarchical setting based on any of these groups.

To begin with modelling, we first ran our data through a number of Glms to get a sense of the important characteristics in squirrel behaviour, as there is little to no background information on the topic. While these results are not deeply studied here, due to our intended focus on Bayesian methods, these preliminary models were essential in pinning down a model that worked after continuous cycles of data re-organization and cleaning.

After this testing, we were left with a few models we wanted to take into stan to explore further and potentially compare their results.

Model Structure

The rough architecture of these models centered around the following theme;

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}\left(\pi_i\right) \\
\pi_i &= \text{logit}^{-1}\left(\beta_0 + \beta_{f,j} \cdot \text{fur color}_i + \beta_{h,k} \cdot \text{highlight color}_i + \beta_{a,l} \cdot \text{activity}_i + \beta_{area,t} \cdot \text{area}_i\right)
\end{aligned}
$$

Where;

$$
\begin{aligned}
&\beta_0 \sim \mathcal{N}(0,1) \\
&\beta_{f,j} \sim \mathcal{N}(0,1) \text{ for: } j = 1,2 \\
&\beta_{h,k} \sim \mathcal{N}(0,1) \text{ for: } k = 1,\ldots,9 \\
&\beta_{a,l} \sim \mathcal{N}(0,1) \text{ for: } l = 1,\ldots 4 \\
&\beta_{area,t} \sim \mathcal{N}(0,1) \text{ for: } j = 1,2,3
\end{aligned}
$$

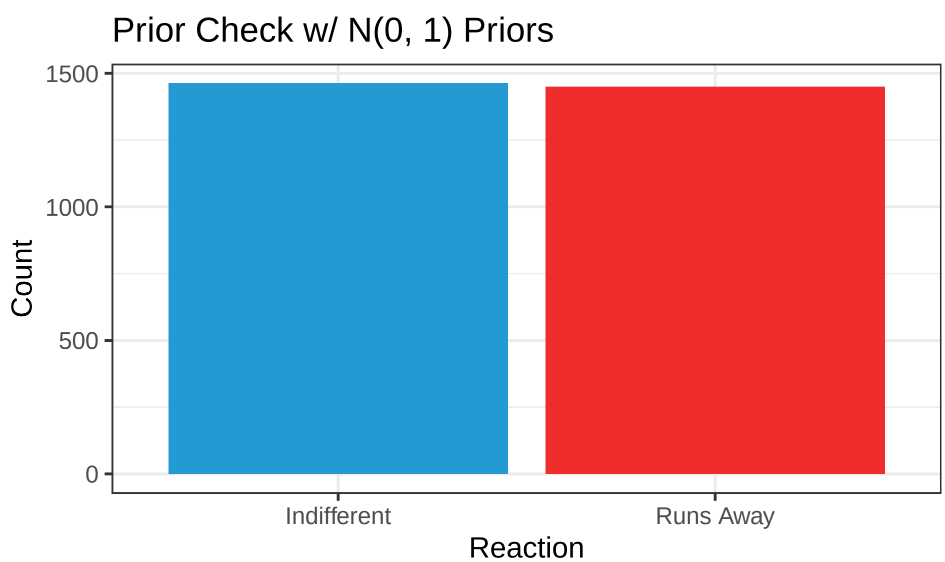
In our first model, we use all of our variables of interest, treating them as covariates to the model, much like how we would fit a typical glm. Here, each β is a vector of category options, leaving out our reference variable, and each β is fit with individual priors. For our priors, we currently are using $\mathcal{N}(0,1)$ priors on all variables. As far as we know, there is almost no valuable information available anywhere that focuses on a topic similar enough to this that could tell us what factors may make a squirrel more or less likely to run. The only information we are currently pulling from is our results in the EDA section of this paper, where we saw extremely subtle relationships between the proportions in each group and variable, but these were never completely clear and free from any contradictions. Thus, we felt it was important to have as weakly informative priors as possible. Furthermore, we believe that any differences that are caused by these factors would likely be relatively small, leaving us fairly confident in these prior selections. However, in the following section will explore these priors a bit more thoroughly.

To change these models, we made a variety of data-reorganization choices, such as grouping activities by their general theme (leisure or active), grouping highlight colors by base colors (gray, black, white, cinnamon), or reducing the number of mixes overall. In turn, however, we found that the best model used the entire pool of options, thus we decided to move forward with this setup, and without reducing the number of groups in each category any more than we already have.

In addition, we also attempted to create hierarchical structures based on the area of the spotting and the census data the observations were coming from, but these models all appeared significantly worse after running, struggling to achieve significance on almost all predictors and intercepts, as well as more frequent convergence issues. As such, we have stuck to fitting our data without any hierarchical aspect for this project, but this stands as a strong 'jumping off' point for future work.

Prior Checks

After defining a rough outline to the model, we must first briefly check our priors and assess how well they are simulating the current data.



As we can see, if we let all of our priors become $\mathcal{N}(0,1)$, then we will estimate an almost equal number of runners and indifferent squirrels, almost every time. We believe that this could be beneficial as it imparts very little of our own beliefs onto the model, but still suggests plausible values for the outcomes, and doesn't heavily bias the results in any way. Thus, we consider these to be our ideal settings for weak information. However, we saw previously that the majority of squirrels in both datasets are actually indifferent to humans' approach. Thus, we may want to skew these results slightly, but that would entail fixing the outcome after observing the data which can be troublesome. Furthermore, the issue here is that if we do this by changing covariate priors we will impart a bunch of information that we do not have, or even see in the data, thus, we are really only left to reducing the intercept term, β_0 to something like $\mathcal{N}(-1,1)$, after which point the data looked much more sensible as we see around twice the number of indifferent squirrels as we do runners. But we do take this option with caution and compare it with our other models to determine if this setting actually helped.

Now that we have briefly checked out prior distributions, we begin with running our models in stan.

(Note, these models take a very long time to run (Around an hour on a strong PC), so for each model we saved their results to an RDS file and uploaded them, in our Rmd file we have also included spaces to read in the data from these RDSs)

Prior Settings In addition to these model variations, we also created a few models that included alternative prior settings. Namely, instead of our $\mathcal{N}(0,1)$ settings, we also created some models that used $\mathcal{N}(0,10)$ priors as well. In doing so, we could better experiment to see if our hypothesis that these effects will be small in magnitude is actually true. In addition, for both of these settings, we also used our prior knowledge by setting our intercept prior at $\mathcal{N}(-1,1)$, to test if these provided stronger predictions. After running these alternative prior settings, however, we found these again produced significantly worse results and missed significance on a greater majority of parameters, thus in our analysis we have stuck with our $\mathcal{N}(0,1)$ and $\mathcal{N}(0,10)$ settings, which will help cut down on space as well.

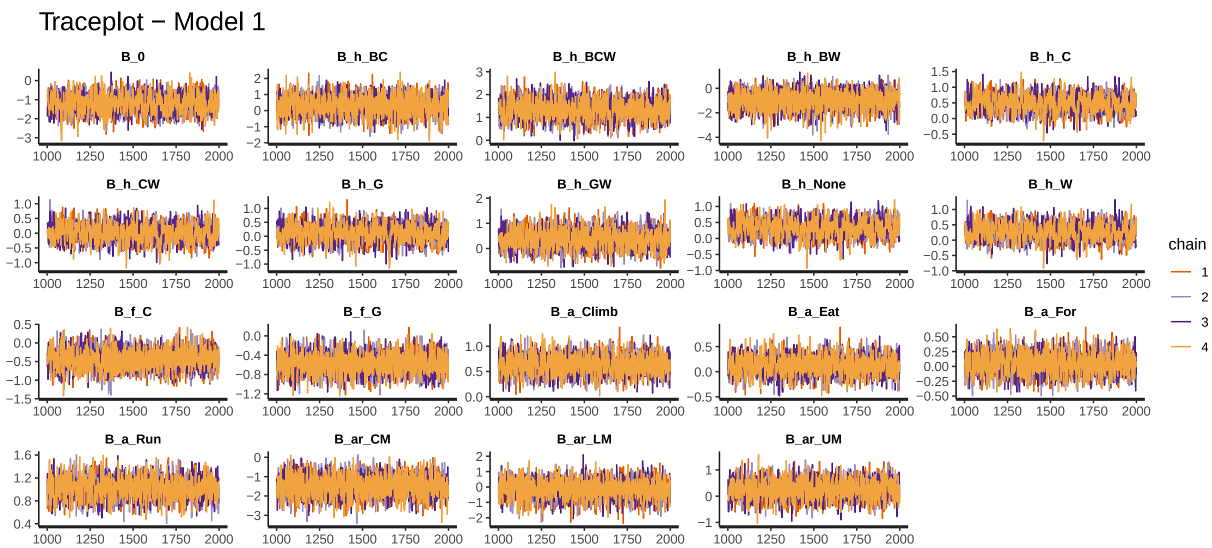
Model Plots and Results

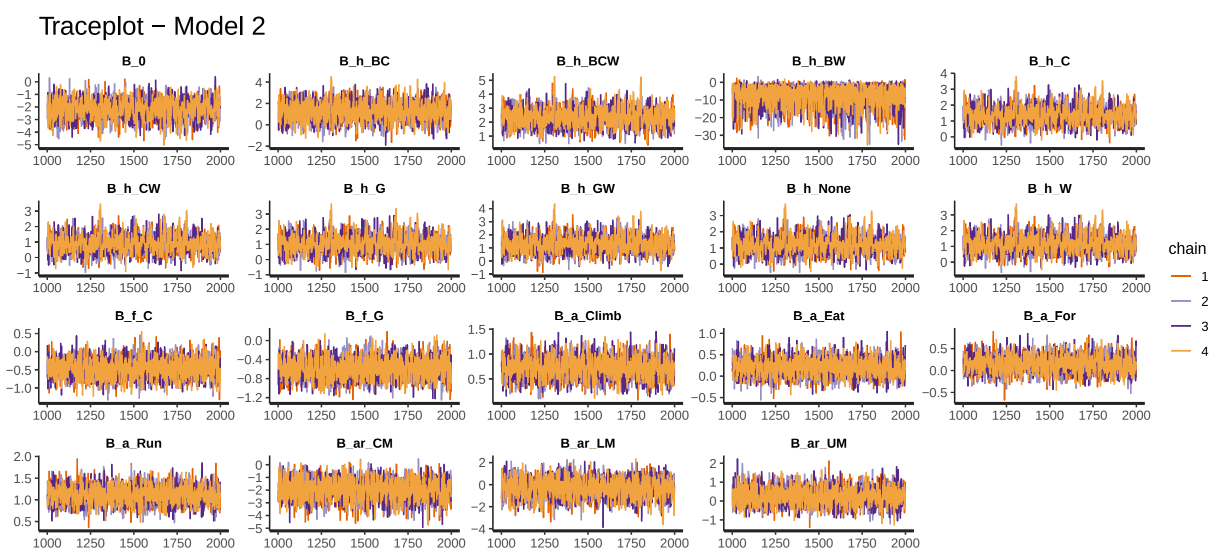
Diagnostics

Before we take a look at the results of our models, we first examine some diagnostic results from our models.

Firstly, while not displayed here, all of the coefficients in all of our models attained Rhats that were within 0.005 of 1.00, making us very confident in the convergence of our model parameters.

We then move to examine the traceplots and pairs plots of our models to gain a more visual understanding.



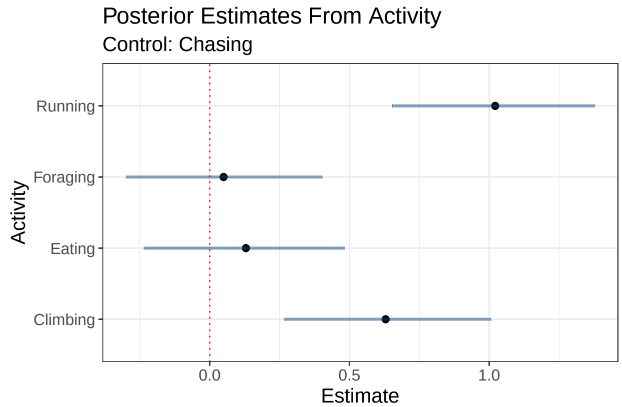


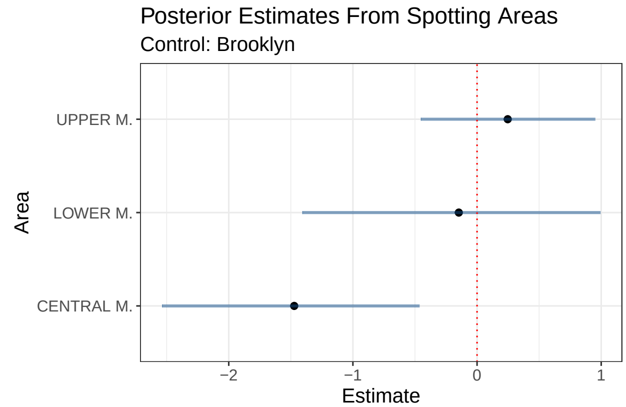
In our traceplots, we can (somewhat) clearly see all chains converging quite well and quite quickly, providing further confidence in the model's overall convergence.

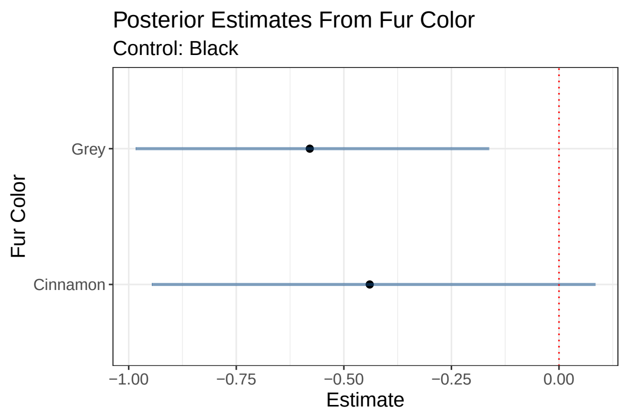
Next, for our pairs plot, due to the high number of included variables in the model, we have omitted these plots. However, within these plots we see very little relationship between any of the groups of covariates. We do see some relationships between covariates and the intercept, but these are fairly light in general, and are likely not going to cause any significant issues. Furthermore, all of our variables are categorical, so there is almost nothing we can do in terms of centering or standardizing these values if we actually wanted to fix the issue.

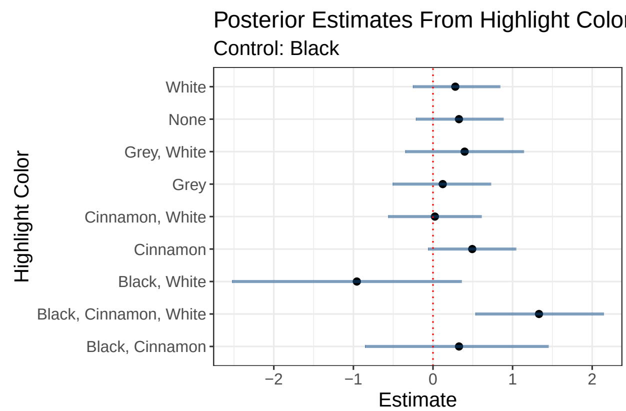
Coefficient Results

To begin with exploring the results of our models, we first assess their coefficient estimates.









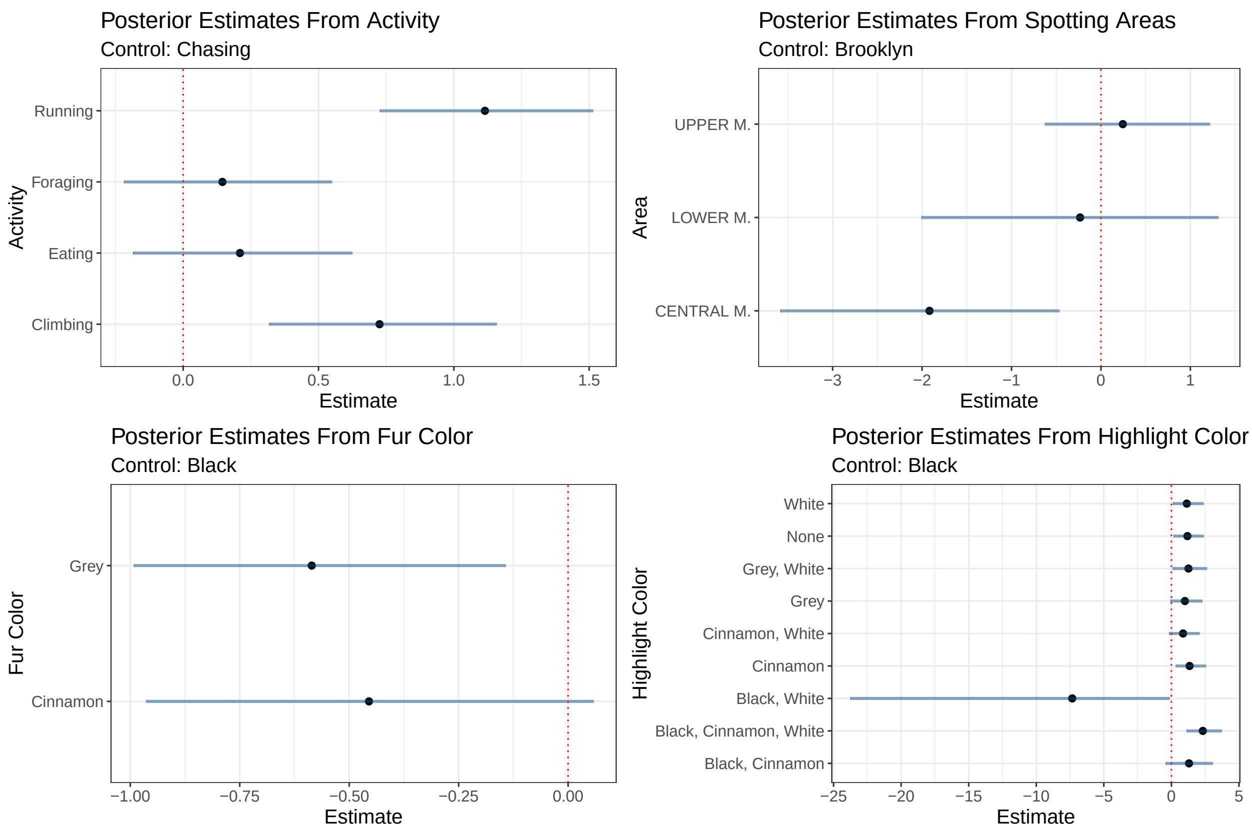
In our first model, with all $\mathcal{N}(0,1)$ priors, we see some interesting trends when plotting our parameter coefficient estimates and 95% credible intervals. Firstly, from our activity variable, we can see that squirrels observed to be climbing or running have higher odds of running when approached by humans, when compared with a squirrel that is chasing something. This makes sense when we consider the nature of each action, as it suggests that squirrels in more agitated states may be more likely to flee from humans, than those engaging in actions that display more comfort. This is mirrored when we compare the significant increase that running has over eating and foraging, two fairly relaxed activities. That being said, we did not see significant results on squirrels currently eating or foraging, noting their CI's overlap with 0.

Moving on to our plot of squirrel primary colors, we can see that compared to black squirrels, grey squirrels show a lower tendency to run which is statistically significant. While we don't see statistical significance in our estimate for cinnamon squirrels, the distribution does strongly suggest that these squirrels are less likely to run as well, but again we cannot conclude this based on the parameters CI. Interestingly, this result somewhat disagrees with our observations from our EDA where we saw that the areas with more cinnamon squirrels actually tended to see more running squirrels. However, this does connect up with our plot of squirrel reaction by fur color as well, where we saw black squirrels showing the greatest proportion of runners in the data.

Next, in our plot of highlight effects, we only saw 1 truly significant effect, and this was from the squirrels with a combination of black, cinnamon, and white highlights, which suggests that these squirrels have a slightly higher tendency to run from a human as compared to a squirrel with only black highlights. The remaining coefficients, however, all showed a lack of significance with CIs overlapping 0. We do see that the Cinnamon highlight group is very close to being statistically significant, where it would suggest that these squirrels have a higher tendency to flee compared to black highlight squirrels as well, however, due to the strictness of the 95% CI we cannot conclude it as significant. This could potentially be a case of an unlucky set of MCMC iterations in stan due to our seed, but re-running it for this sake would be somewhat dishonest.

Finally, in our plot for area (top right), our model suggests that, compared to Brooklyn, squirrels in Central Manhattan have a significantly lower tendency to flee when approached. This result matches up quite well with our previous observation in our EDA where Central Manhattan had a much lower proportion of fleeing squirrels than any of the other areas included. The remaining areas, however, do not show any significant differences, but this is not incredibly surprising as these groups appeared quite similar when plotting previously.

Next, we move on to examining our model that used all $\mathcal{N}(0, 10)$ priors.



First, in our estimates for squirrel activity, we see almost identical results to our $\mathcal{N}(0, 1)$ model, with significant positive effects on climbing and running as compared to chasing, but no big difference seen on eating and foraging.

Next, when we move to our estimates of base fur color, we see an almost identical result to our previous model again, with grey squirrels showing a lower tendency to flee than black squirrels, and cinnamon squirrels showing a lower tendency in most cases, but still lying just outside of the bounds for statistical significance.

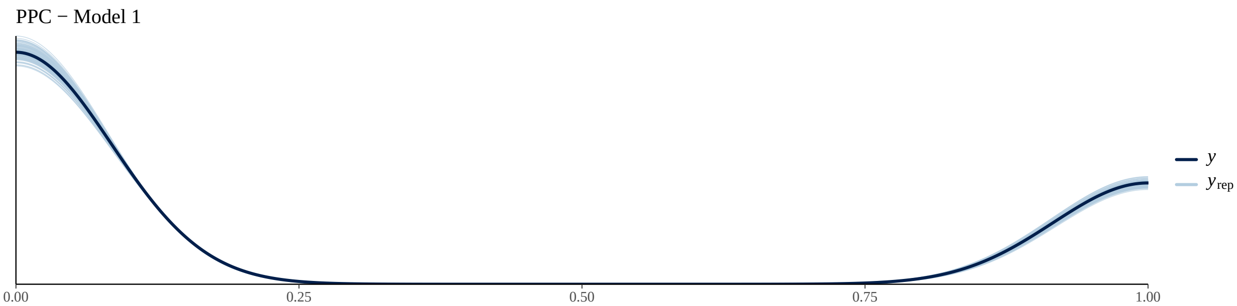
In our highlight plot, however, we see much greater levels of significance. Again we are comparing all estimates to the control category, which in this case is black highlights. Firstly, we can see that squirrels with (black, cinnamon, and white) highlights have a higher tendency to flee just as we saw before. We can also see a slightly higher tendency to run for squirrels with highlights that are; (cinnamon), (grey and white), (non-highlighted), and (white). Furthermore, we can see that Grey and cinnamon-white highlighted squirrels have a majority of their distribution suggesting they are more likely to run, however, these estimates' CIs run just over 0, rendering us unable to conclude these effects. Finally, and conversely, we can see that squirrels with a black and white highlight have a lower tendency to run than squirrels with just black highlights.

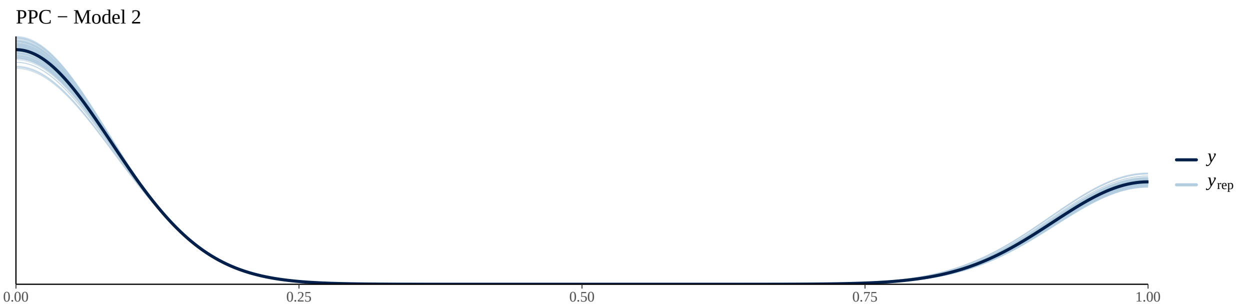
Finally, in our plot of area-based estimates, we don't see any significant differences from our $\mathcal{N}(0, 1)$ model again, as we note that squirrels in Central Manhattan appear to have a lower tendency to run than squirrels in Central Manhattan. And again, we fail to see significant results coming from Lower and Upper Manhattan.

PPCs

Now that we have explored our models' coefficients, we move on to assessing some posterior predictive checks, as we wind down to selecting our final model.

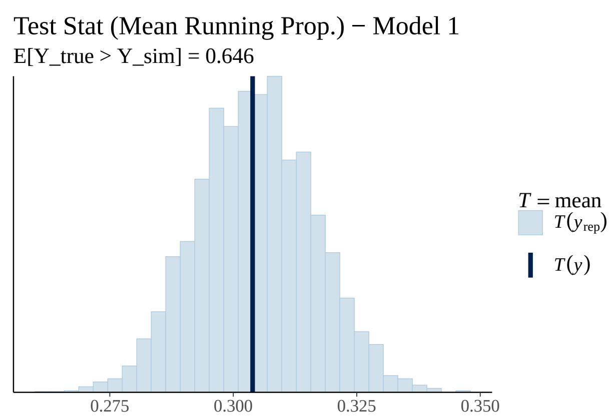
To do so we will be using three primary methods, PPC plots, test statistics, and confusion matrices.

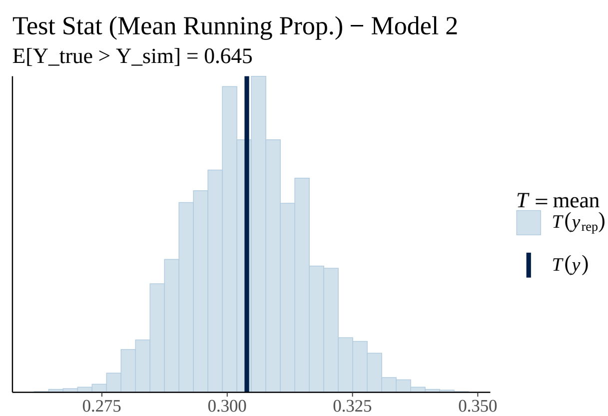




Firstly, from our PPC plots, we can see that both models do quite a good job of fitting the true data, as there are almost no simulated lines that diverge in either plot. However, we do notice a slightly wider spread of points at the ends of the plot in our $\mathcal{N}(0, 10)$ model, suggesting that our first model may be doing a slightly better job of simulating our data. There is little other information that can be derived from this plot, however, as the shape is heavily restricted due to the nature of logistic regression. In turn, our test statistics and confusion matrices become more helpful for discerning how well each model can predict data.

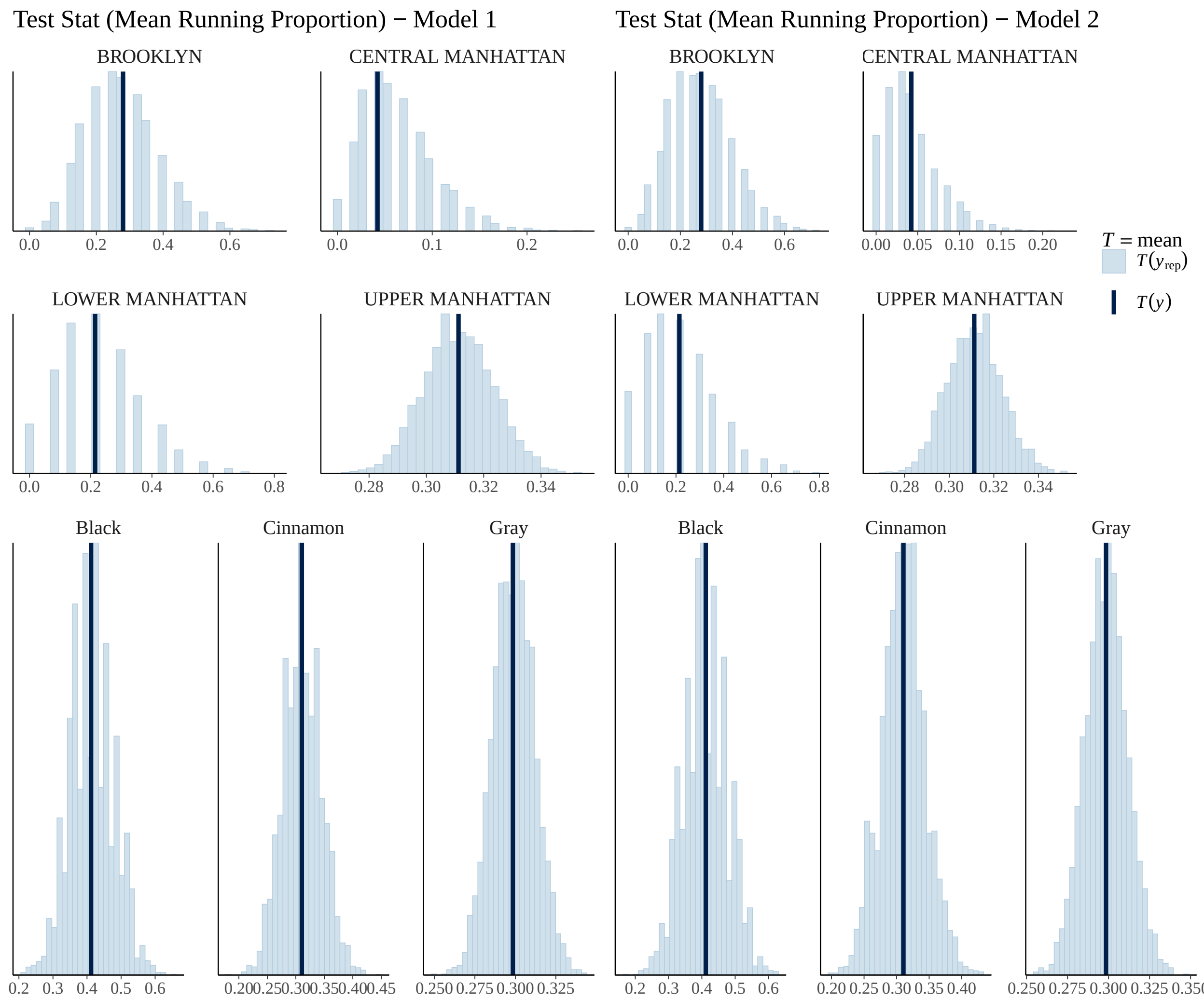
Next, we can take a look at some of our test statistic plots. As a note here, our data is filled entirely with categorical variables, and our outcome is binary, which heavily restricts the freedom we have with choosing a test statistic as we are largely tied to using the mean and conditioning on each variable.





To start, we examined the mean response to humans across the entire population, and compared it to our simulated populations for each model. As we can see, both models do a very good job and fit the true value nearly dead center at the peak of their distributions. Furthermore, we can calculate the proportion of values above and below the truth to pin the differences down numerically. In doing so, we find that both the $\mathcal{N}(0,1)$ and $\mathcal{N}(0,10)$ predict datasets with means below the true value 65% of the time, which is fairly close to our ideal goal of 50%. These results make us fairly confident that both models are able to capture some of the basic aspects of the data.

Next, we move on to plotting these relationships by group.



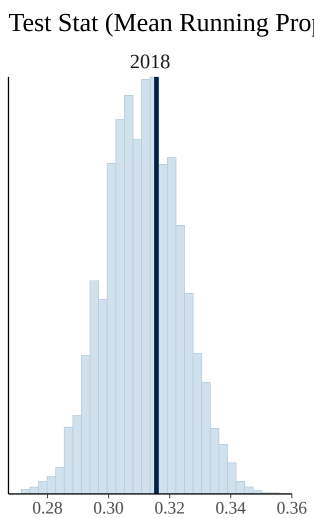
For our first groups, we plot these test statistics by spotting area. As we can see, within both groups we again have a very good fit of the distribution around the true value, placing it at or very near the areas of the highest simulation proportion, providing further confidence in our models' abilities to predict new data. Then when we move to fur color, in our bottom row, we see very good fits again across both models and all covariates, another very good sign of our models' strength. Thus, overall we remain fairly confident that our model is performing quite well when it comes to predicting new data.

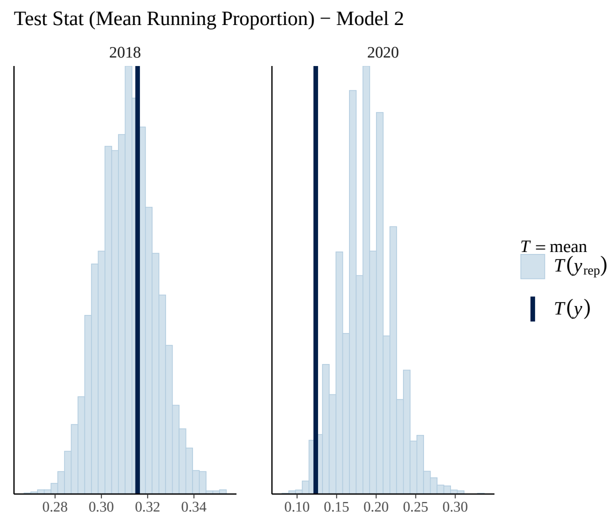
Table 3: Confusion Matrix - Model 1

	Pred +ve	Pred -ve
True +ve	0.3413	0.6587
True -ve	0.2888	0.7112

Table 4: Confusion Matrix - Model 2

	Pred +ve	Pred -ve
True +ve	0.3440	0.6560
True -ve	0.2876	0.7124





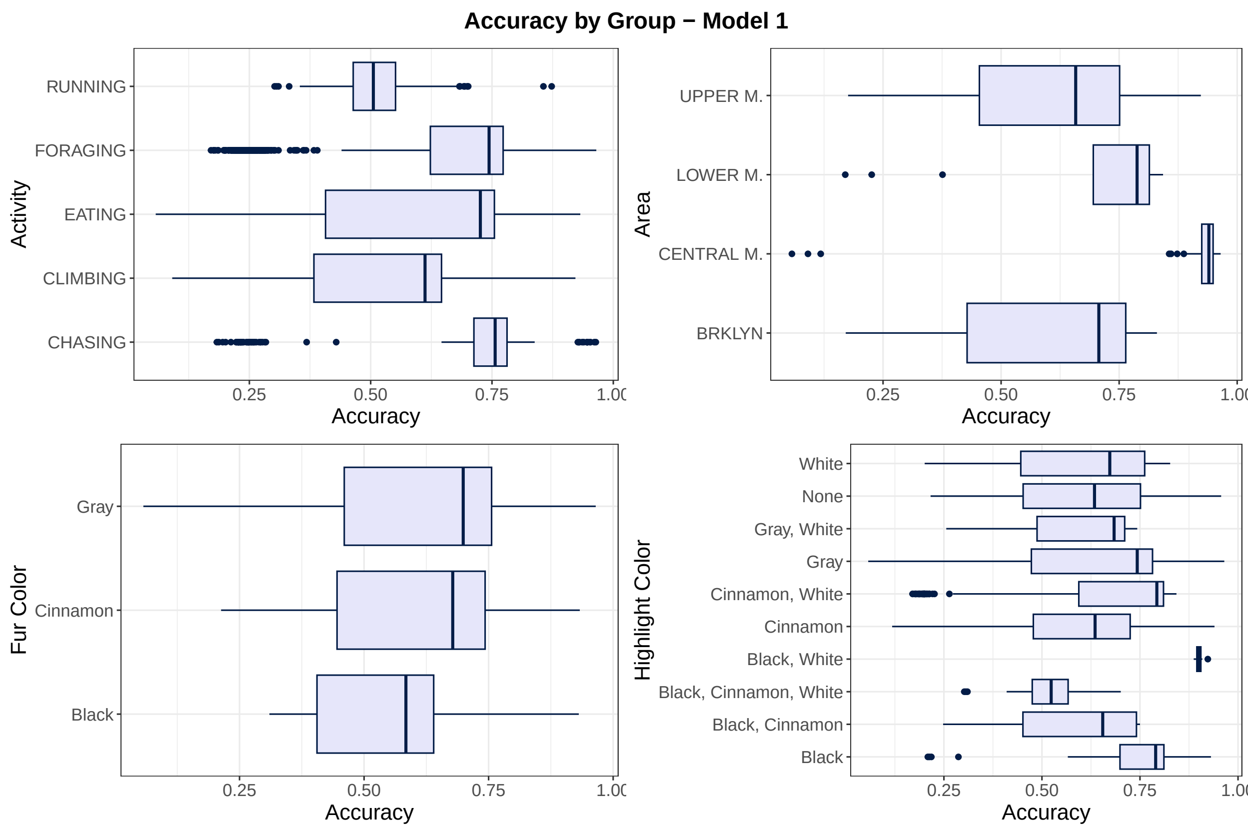
For a final plot, we examined the posterior distribution and plotted our test statistic against the census data that the observation came from. Here, we can clearly see that the 2018 data does a much better job of fitting the true value in the middle of the distribution. Conversely, however, the 2020 dataset shows a true value well below most of our simulations, leading us to believe that we may not be able to accurately simulate data from this census. This leads us to wonder what went wrong previously when we tried to fit a model hierarchically on the data, and whether it would have been a better idea to focus on just one dataset. However, that was not our chosen direction and would require a re-haul of much of this paper, thus, we make note of this result to discuss later with our other limitations and plans for future work.

As our final method for posterior checks, we examine the accuracy of our models. First, we have evaluated the TPR and FPR rates from our data, to create a confusion matrix for each model so that we can assess their individual accuracy for classification. To do so, we simply compare the true outcomes with each simulated outcome run, and evaluated over each data point and over each iteration. Firstly, between our two models, we see almost identical confusion matrices, leading us to believe that both models are of relatively equal strength when it comes to predicting outcomes. We do, however, see that our true positive rate is quite low, at just around 34%, which may be significant of the small proportion of squirrels we saw running in the first place giving trouble to our model when it comes to predicting these outcomes.

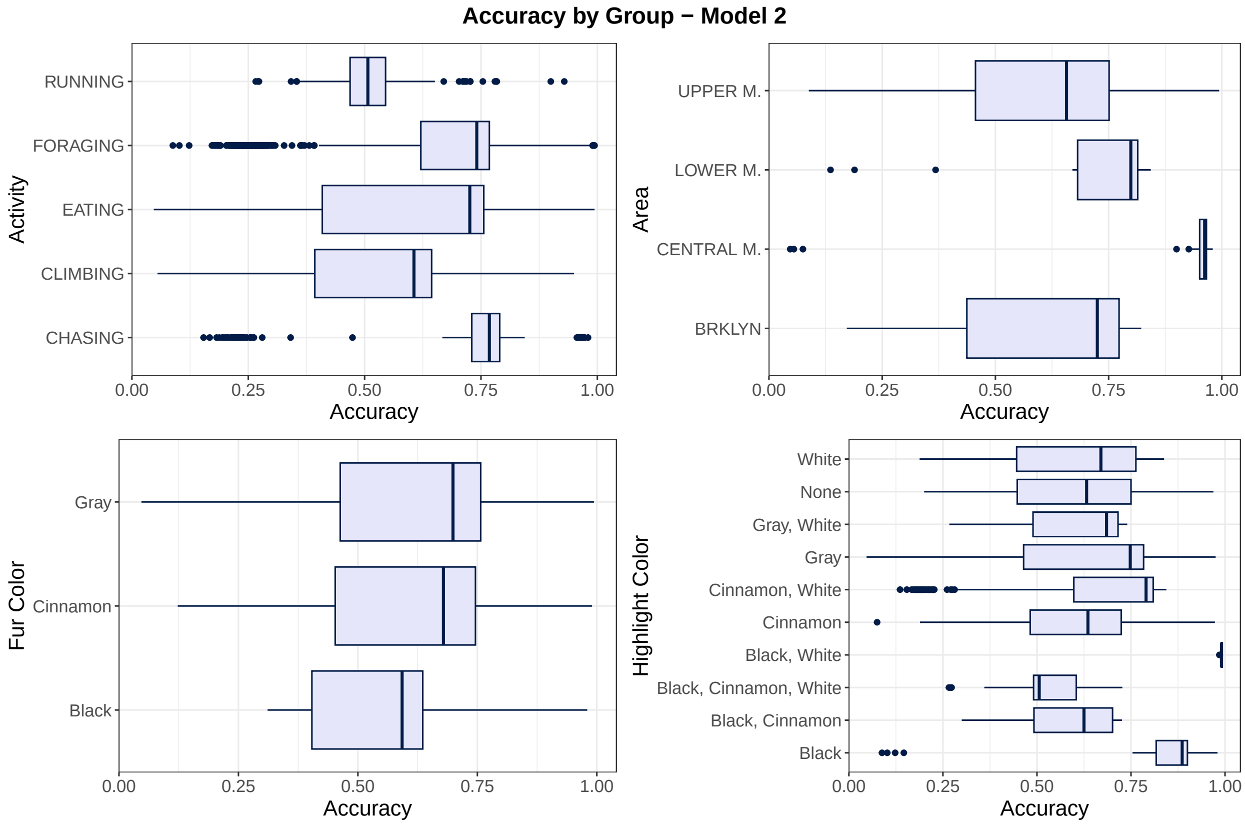
We can also examine the overall accuracy of both models.

	Model 1	Model 2
Accuracy	0.5988	0.6005

Overall, in both models we see an accuracy of around 60% in the first 1000 runs. (Only ran first 1000, running the full set would crash the session):)



To further understand our accuracy across different settings, we have created boxplots of our accuracy by categorical groups. Briefly, within the activity plots, we can see that we tend to see the best accuracy in squirrels currently chasing something, and the worst accuracy in squirrels currently running, with the other factors remaining fairly close together. With fur color we remain fairly similar across groups, achieving accuracies between 55 and 75%. Next, within highlight colors, we again see a very close distribution of accuracy over almost all the categories except the black and white highlights, where we appear to have significantly higher accuracy. Finally, we can see that within the area categories, our model does the best job at predicting the reactions of squirrels in Central Manhattan, which may be a result of them having the lowest proportion of runners as we note the low TPR from this model.



Now, observing these results from our second model, we find very similar outcomes. Squirrel area has very small differences between groups, tending to stay between 50 and 75% accuracy. Fur color again appears to share similar estimates across all categories, between 55 and 75%. Then, within the highlight categories, we see similar results again, except for a slightly higher accuracy, sitting above 80% for squirrels with black highlights, and an almost perfect accuracy for squirrels with black and white highlights. Lastly, in our plot against spotting area, we found our model to report relatively similar accuracies across all groups except Central Manhattan again, which we believe is due to their lack of fleeing squirrels.

Model	Classification.Accuracy
1	0.6988
2	0.7012

As we saw, the accuracy across both models is not great overall, however, if we instead measure our mean predicted outcome, that is, by taking the mean across each row of our predicted values, we can set a bound of 0.5 and indicate our overall predictions as 1 if the mean estimate is above 0.5, and 0 if it is below, similar to a usual classification setting. In doing so, we find that our model's accuracy on the data itself is closer to 70%, which is considerably stronger. This may be a further result of the lack of squirrels actually running in the data, and may point to our model again struggling with these observations. This does also invite the possibility of tuning this threshold to find the most optimal point for classification.

Model Comparison

Lastly, before we interpret our results, we must first make some final comparisons between the models to select a winner. To do so we will compare the models based on their loo ELPD values.

As we can see, in comparing our two models, their ELPDs are almost identical, separated only by 0.3. As such, choosing a model is left mostly up to our discretion. In turn, due to the greater number of significant covariates, and marginally higher accuracy, we have selected our second model, which uses the $\mathcal{N}(0, 10)$ priors, as our final model.

Table 5: Model 1 Loo Estimates

	Estimate	SE
elpd_loo	-1720.6975	22.7355
p_loo	15.4665	0.6747
looic	3441.3951	45.4709

Table 6: Model 2 Loo Estimates

	Estimate	SE
elpd_loo	-1721.0150	23.2424
p_loo	18.5343	1.1886
looic	3442.0300	46.4848

Final Model Resutls & Interpretations

Now that we have chosen our final model, we can begin interpreting the results of our significant effects, remembering here that our estimates are on the log-odds scale, and must be transformed through exponentiation to more easily interpret.

As we saw previously, we have significance on just part of this model.

Firstly, we see an intercept value of -2.16, which, after exponentiating, suggests that a baseline squirrel, that is, one with black fur and black highlights, currently chasing something in Brooklyn, has an odds of running of about 0.12. Within the highlight category, we get an estimate for squirrels with black, cinnamon, and white highlights of 2.35, thus, in comparison to squirrels with black highlights, our model predicts these squirrels to have an odds of running that is over 10 times greater. Next, for black and white highlight squirrels, with an estimate of -8.62, our model predicts that these squirrels will have an odds of running that's over 5000 times lower than black squirrels. Cinnamon and grey + white highlights also both showed significant increases in odds of running as we look to their estimates in the data as 1.36 and 1.29, respectively, which suggests an increase in the odds of about 3.9 or 3.6 multiplicatively. We also noticed significance in squirrels with no highlight and those with white highlights, which again both suggested very similar trends, with estimates of 1.20 and 1.15, respectively. Thus, these squirrels are predicted to have an odds of running that is 3.3 and 3.2 times higher than those with black highlights.

Next, when we move to fur color, we only get significance on squirrels that have grey fur, where we get an estimate of -0.58, suggesting that grey squirrels have an odds of running that's about half of what we estimate for black squirrels.

When we look to our activity estimates, we can see that squirrels currently climbing and running, with estimates of 0.73 and 1.11, both are significantly more likely to flee from a human than squirrels currently chasing something. For climbers, we see they have about double the odds of running, while runners have over triple the odds of running when approached.

Lastly, for our area parameter, we only found significance in our estimate on Central Manhattan, where we can see an estimate of -1.95, which suggests that squirrels in this area have about 15% of the odds of running away as squirrels do in Brooklyn.

Table 7: Model Loo Comparison

	elpd_diff	se_diff
model1	0.0000000	0.000000
model2	-0.3174973	2.280615

Table 8: Final Model Interpretations

	mean	odds	sd	min	max	Rhat	significant
B_0	-2.1588	0.1155	0.7792	-3.7288	-0.7146	0.9999	1
B_h_BC	1.3014	3.6743	0.8766	-0.4237	3.0480	1.0006	0
B_h_BCW	2.3539	10.5270	0.6530	1.1220	3.7082	1.0029	1
B_h_BW	-8.6221	0.0002	6.2570	-23.7647	-0.1435	1.0011	1
B_h_C	1.3606	3.8985	0.5652	0.3213	2.5409	1.0022	1
B_h_CW	0.8814	2.4142	0.5767	-0.1807	2.0681	1.0027	0
B_h_G	1.0080	2.7402	0.5848	-0.0776	2.2656	1.0022	0
B_h_GW	1.2854	3.6163	0.6282	0.0915	2.6131	1.0011	1
B_h_None	1.1975	3.3118	0.5616	0.1583	2.3670	1.0022	1
B_h_W	1.1515	3.1629	0.5664	0.1106	2.3590	1.0024	1
B_f_C	-0.4546	0.6347	0.2622	-0.9637	0.0577	1.0008	0
B_f_G	-0.5820	0.5588	0.2159	-0.9920	-0.1432	1.0010	1
B_a_Climb	0.7263	2.0675	0.2088	0.3170	1.1578	1.0002	1
B_a_Eat	0.2105	1.2343	0.2066	-0.1855	0.6238	1.0001	0
B_a_For	0.1513	1.1633	0.1961	-0.2190	0.5488	1.0009	0
B_a_Run	1.1141	3.0469	0.2032	0.7262	1.5137	1.0004	1
B_ar_CM	-1.9498	0.1423	0.7944	-3.5855	-0.4659	1.0000	1
B_ar_LM	-0.2580	0.7726	0.8455	-2.0071	1.3110	1.0010	0
B_ar_UM	0.2556	1.2913	0.4651	-0.6259	1.2157	1.0012	0

Conclusions

In this analysis, we aimed to understand the characteristics and habits of squirrels that make them more or less likely to flee from humans. As we saw, a large portion of this process was organizing the data, due to the consistent amount of non-standardness in the data. However, after this task, we were able to get some models running and see some interesting results.

Primarily, we saw that most of the squirrels in New York are actually more likely to be indifferent when approached by a human. There are some characteristics, however, that do help explain why the ones that do run, run. Firstly, we found fur color to have a relatively small effect, but one that in general suggested that black squirrels are the most likely to flee, whereas grey or cinnamon squirrels have a slightly lower tendency to run. Within the highlights section, this was quite a mixed bag, as many predictors were just on the edge of 0, making us not totally confident in their significance. Regardless, however, this did suggest to us that almost all colorings, aside from those with black and white highlights, had higher tendencies to flee than squirrels with black highlights. For our activity measurement, we saw fairly intuitive results, as those currently in very active states, like running or climbing, are predicted to be slightly more likely to flee, similar to what we observed before. Lastly, for area, we found that squirrels coming out of Central Manhattan had a much lower tendency to flee as compared to those from Brooklyn. We also saw Central Manhattan with the highest proportions of squirrels appearing unfazed to human contact, potentially adding strength to our argument of squirrel migration over time.

While these results seemed rather strong to us, when we pulled the models in for deeper inspection and accuracy testing, we noticed that our model struggled with its True Positive Rate (TPR), indicating a significant fault in the model. Thus, we have found that predicting the exact behaviour of a squirrel is extremely hard based on the limited amount of data we had access to. While we still get some interesting results and trends, there is significant work that needs to be done in further model examination to build a stronger and more accurate model of these characteristics and traits.

Limitations and Future Work

Our primary limitation throughout this work was through the use of stan, as we were relatively bottlenecked by the time it took to run each model. This was primarily a result of a combination between a large dataset, and a fairly high number of parameters that needed estimation. The issue here, however, was that upon sampling the data, we begin to quickly lose significance on our predictors as well as overall strength to the model, thus we were relatively stuck with this as our only viable option. Furthermore, we were also somewhat limited by our options for data to include in our model. As this was made to be easily collectible by the volunteers, there were very few columns of any highly detailed features of the squirrels. Again, however, there is little that we can do to alleviate this, aside from attempting to create our own values based off things like the general trend in an activity, or by using interaction terms between our fur and highlight colours. While we attempted this already, and found no immediate success, our first success was found through a similar process, leading us to believe that additional work can be done in the way of re-organizing the data to potentially achieve better results. Moreover, we still believe there is quite a bit of promise in designing a model that uses a hierarchical structure to answer this question, but it is likely that this will take considerable effort both in choosing a variable and organizing the data accordingly.

Regardless of these drawbacks, however, we remain fairly satisfied with our model, as it provides some very interesting information on a topic that we have never really seen before, and gives somewhat of a starting point for future investigations into the habits and characteristics of common tree squirrels.

Bibliography

1. Cooper, Christopher A., Allison J. Neff, David P. Poon, and Geoffrey R. Smith. "Behavioral Responses of Eastern Gray Squirrels in Suburban Habitats Differing in Human Activity Levels." Northeastern Naturalist 15, no. 4 (2008): 619–25. http://www.jstor.org/stable/25177145.

2. Nowak, T., Riahi, S., and Ward, M. "Shyness and boldness in squirrels: risk-taking while foraging depends on habitat type" Lake Forest College News (2018). https://www.lakeforest.edu/news/shyness-and-boldness-in-squirrels-risk-taking-while-foraging-depends-on-habitat-type

3. Rademacher, E. "Responses of urban gray squirrels (Scurius carolinensis) to humans and conspecifics in an area of Boston Common" Boston University Arts & Sciences Writing Program, (2018). https://www.bu.edu/writingprogram/journal/past-issues/issue-11/rademacher/rademacher-writer/

Data

2018 Census:

- rfordatascience, tidytuesday (2019). https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-10-29

2020 Census:

- The Squirrel Census, (2020). https://www.thesquirrelcensus.com/data