Various factors associated with Eastern gray squirrels' interactions with humans in Central Park

AUTHOR
Raima Saha

Abstract

The Eastern gray squirrels were introduced to New York City in the 1870s (Allen, "Getting to Know"). Though they've lived there for a while, little is known about them, especially on how they interact with humans. With Principal Component Analysis, the Chi-Squared Test of Independence, and the K-Nearest Neighbors algorithm, it will be determined what factors, such as Geographical Location, Location, Shift, and Primary Fur Color, impact a Central Park squirrel's interaction with humans. It was found that Geographical Location, Location, Shift, and Primary Fur Color were all significantly correlated to Human Interaction. Combined with other data about the Eastern gray squirrel from outside New York City, evolutionary patterns can be found to better understand evolution as a whole and the psyche of the Eastern gray squirrel so they can continue coexisting peacefully with humans in New York City.

Background

The Eastern gray squirrel was first introduced to New York City as part of an attempt to beautify and naturalize what was becoming a very crowded and polluted city. Over a century later, this species has not only integrated seamlessly to city life but as also learned to peacefully coexist with humans. Squirrels are a crucial part of the ecosystem, as they are the food source for many carnivores and bury acorns that later become trees (Allen, "Getting to Know"). They remain an integral part of New York City, whose population has grown to over 8 million people (2020 Decennial Census).

This led me to wonder: What factors, such as Geographical Location, Location, Shift, and Primary Fur Color, impact a squirrel's behavior toward humans?

Analyzing squirrel's interactions with humans could lead to improved relationships between the two populations so that they can continue coexisting peacefully and maintain a healthy equilibrium.

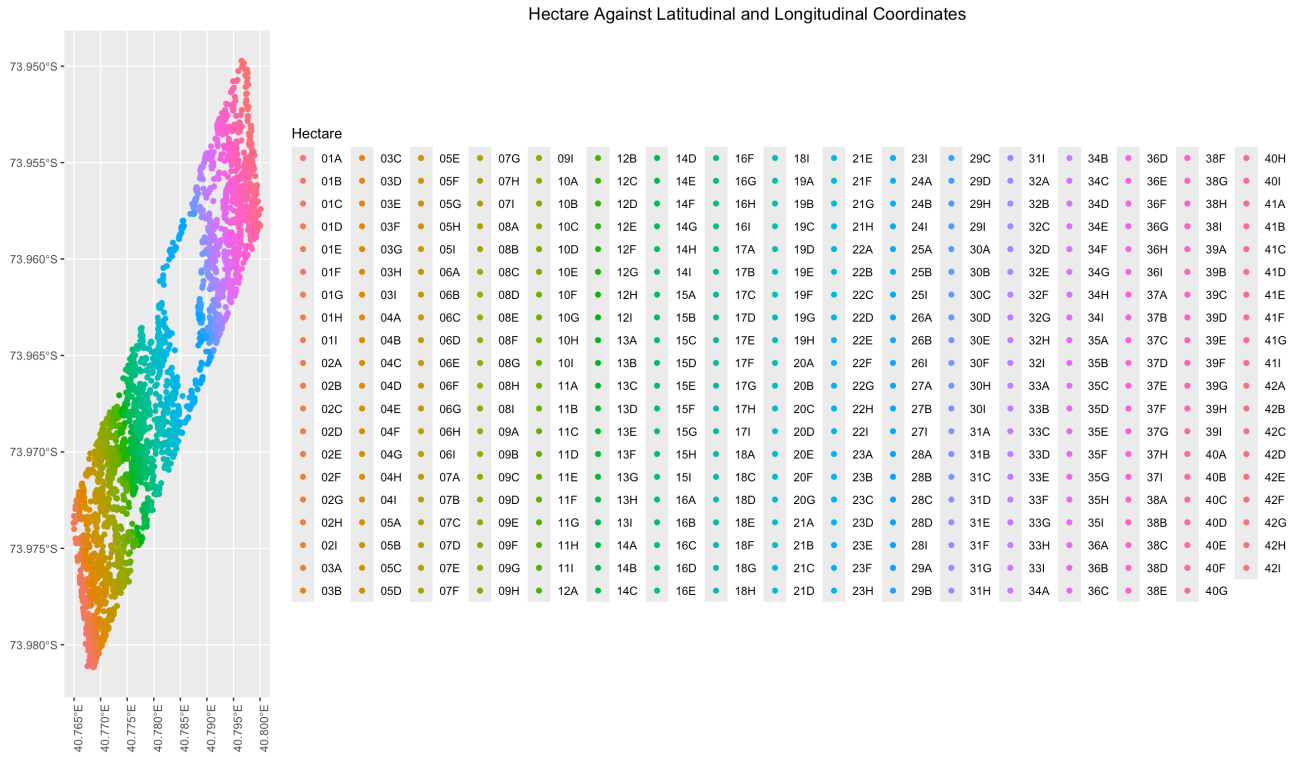
The dataset used for this analysis is from The Squirrel Census, a multimedia science project dedicated to the Eastern gray squirrels in New York (Allen, "2018"). The dataset, 2018 Central Park Squirrel Census, contains 3,023 observation and 31 variables detailing the age, fur color, activity, and behavior of each of the squirrels.

Approach

I focused on the relationship between Geographical Location and Human Interaction, Location and Human Interaction, Shift and Human Interaction, and Primary Fur Color and Human Interaction.

It is important to note that I converted many of these categorical variables into numeric types using various encoding methods, the details of which will be included in each of the descriptions of the variables.

Geographical location: This is where the squirrel is found. For parts of the analysis that require numerical data, I used the `latitude` and `longitude` variables to conduct such analysis. However, some parts of the analysis, such as the Chi-Squared Test of Independence, required categorical data, in which case I used the `hectare` variable (manipulated differently than the original provided variable). This variable divides the area of Central park in a grid-like formation and is labeled 01A to 42I. An example of how it is visualized is provided below (Figure 1). For the sake of consistency, I will refer to all these variables as "Geographical Location" as part of the larger context, but will explicitly refer to them by their given variable names whilst explaining the analysis.



Location: `location` describes the location of the squirrel relative to the ground. “Ground Plane” means the squirrel was found on the ground and “Above Ground” means the squirrels was found above the ground. This variable was numerically encoded in `location_num` so that “Ground Plane” equaled 0 and “Above Ground” equaled 1.

Shift: `shift` describes the time of day the squirrel observation was recorded, “AM” or “PM.” This variable was numerically encoded in `shift_num` so that “AM” equaled 0 and “PM” equaled 1.

Primary Fur Color: `primary_fur_color` describes the main color of the squirrel, either “Gray,” “Cinnamon,” or “Black.” The dataset also contains the variable `highlight_fur_color`, so it is important to make this distinction between the two fur colors. This variable was numerically encoded in `primary_fur_color_num` so that “Gray” equaled 1, “Cinnamon” equaled 2, and “Black” equaled 3. A value of 0 indicates an NA value.

Human Interaction: This variable was initially separated so that each observation had a variable for all three types of interaction (`approaches`, `indifferent`, `runs_from`) and TRUE and FALSE to indicate which one the squirrel exhibited. I created a new categorical variable `human_interaction` that had “approaches,” “indifferent,” and “runs from” as different variable options. Additionally, I also numerically encoded this variable as `human_interaction_num` so that “approaches” equaled 1, “indifferent” equaled 2, and “runs from” equaled 3. A value of 0 indicates an NA value and that no human interaction was recorded.

My approach to my research question can be broken into the following parts:

1. Principal Component Analysis
2. Chi-Squared Test of Independence
3. K-Nearest Neighbors Model

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset. In other words, it determines which features, or variables, in the dataset account for the most variance and are most important. Considering that there are 31 variables in this dataset, I wanted to see which factors I was most interested (Geographical Location, Location, Shift, and Primary Fur Color) accounted for most of the spread of the data and were most impactful to human interaction. Since PCA can only be done on numeric variables, I used the encoded versions of the categorical variables to conduct this analysis.

Chi-Squared Test of Independence

The Chi-Squared Test of Independence is a statistical analysis test used to see if there is an association between two categorical variables. In other words, it describes whether two variables are related or not. I used this test to see if Geographical Location, Location, Shift, and Primary Fur Color were significantly associated with Human Interaction, using the categorical versions of each variable. My null hypothesis was that there was there was no significant association between Geographical Location, Location, Shift, and Primary Fur Color and Human Interaction.

My alternative hypothesis was that there was a significant association between Geographical Location, Location, Shift, and Primary Fur Color and Human Interaction. I set my significance threshold at 0.05. This means that if the p-value returned from the analysis is less than 0.05, the variables I would be comparing would be significantly related and I would reject my null hypothesis. Otherwise, if the p-value is greater than 0.05, the variables I would be comparing would not be significantly related and I would fail to reject my null hypothesis.

K-Nearest Neighbors (KNN) Model

K-Nearest Neighbors (KNN) is a supervised machine learning learning model to classify data into certain groups. Splitting the data into training and testing data, I used this algorithm to create a model to see if it was possible to accurately predict how a squirrel would act based on Geographical Location, Location, Shift, and Primary Fur Color.

Results

Principal Component Analysis (PCA)

Table 1 displays a covariance matrix yielded after conducting PCA on the variables `latitude`, `longitude`, `primary_fur_color_num`, `shift_num`, `location_num`, and `human_interaction_num` Positive values indicate a positive relationship between certain principal components, while negative values indicate a negative relationship between certain principal components. However, as a table, it is hard to come to any conclusions.

	latitude	longitude	primary_fur_color_num
latitude	1.00000000	0.907417636	-0.0232677507
longitude	0.90741764	1.000000000	-0.0473338546
primary_fur_color_num	-0.02326775	-0.047333855	1.0000000000
shift_num	0.04588772	0.027218682	-0.0264493455
location_num	0.01208617	-0.006981278	0.0002225279
human_interaction_num	0.16733918	0.189370646	-0.0098380451

	shift_num	location_num	human_interaction_num
latitude	0.04588772	0.0120861718	0.167339176
longitude	0.02721868	-0.0069812785	0.189370646
primary_fur_color_num	-0.02644935	0.0002225279	-0.009838045
shift_num	1.00000000	-0.1024105708	0.040410435
location_num	-0.10241057	1.0000000000	0.047885258
human_interaction_num	0.04041044	0.0478852577	1.000000000

Figure 2 returns a correlogram visualizing the principal components in the covariance matrix. Looking at the output, `latitude` and `longitude` seem to have a strong positive correlation, while the other variables don't seem to be correlated.

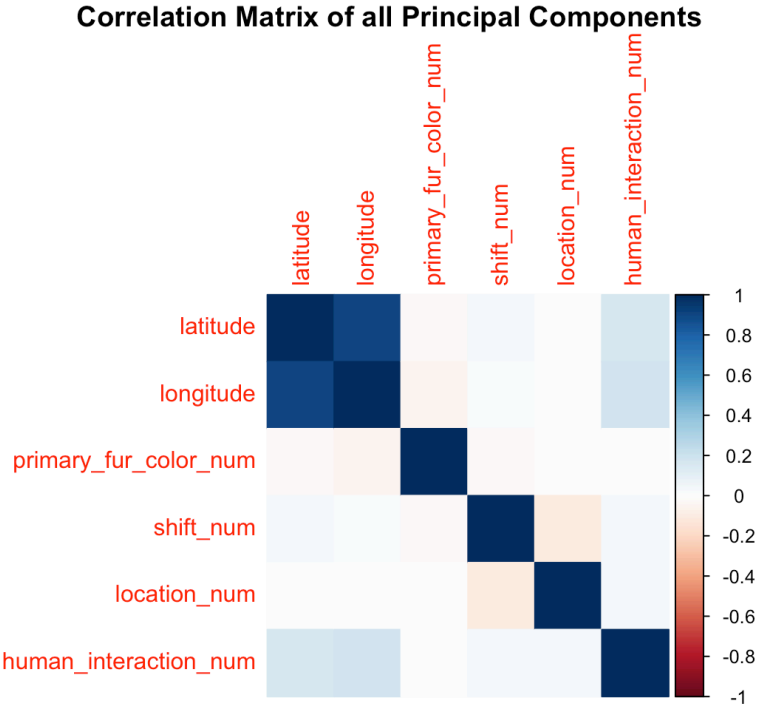


Table 2 shows the total explained variance. It is said that you need the number of components that can explain up to or more than 0.9 of the variance. Looking at the table, the first five components explain more than 0.9 of the variance and by this proportion are the most important and should be the ones considered.

	number_of_components	total_explained_variance
1	1	0.3299030
2	2	0.5140995
3	3	0.6805016
4	4	0.8414838
5	5	0.9847537
6	6	1.0000000

Figure 3 visualizes the explained variance ratio and the cumulative explained variance values from Table 2.

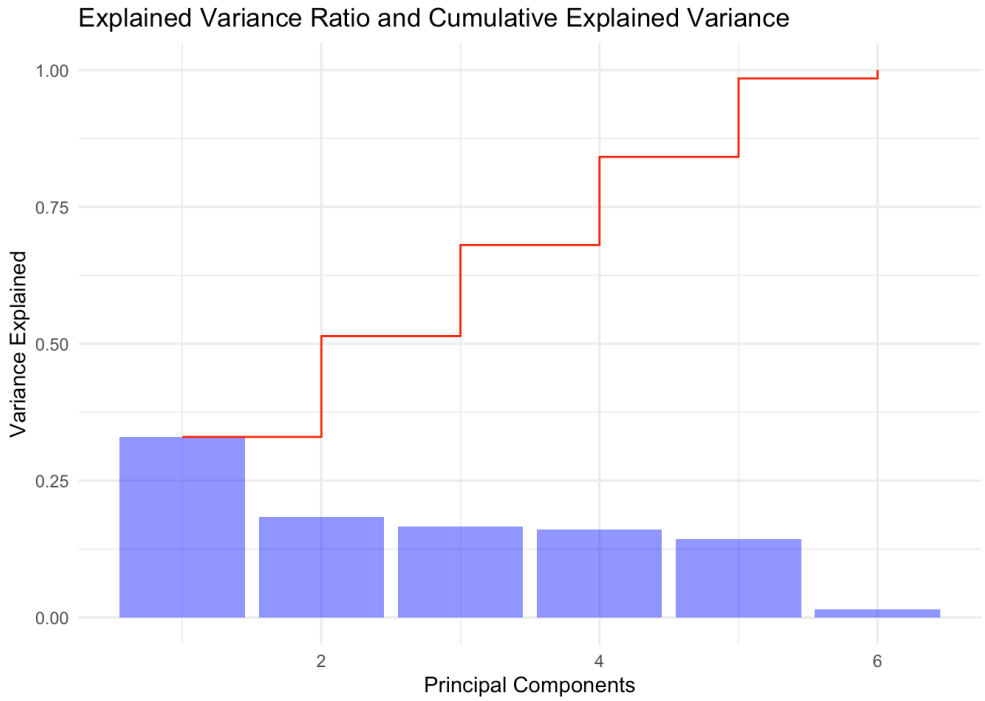


Figure 4 shows a Scree plot of all five principal components. It is consistent with the Explained Variance Ratio in showing that 5 components is enough to explain majority of the variance within the chosen variables.

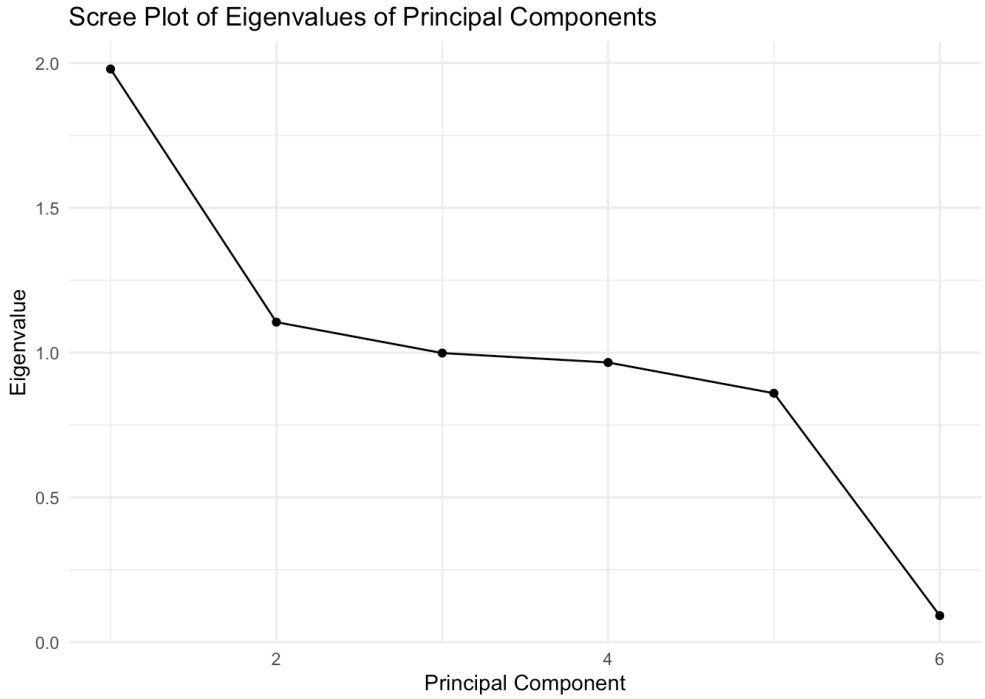


Figure 5 shows a 2D representation of principal component 1 against principal component 2. The points are colored based on Human Interaction.

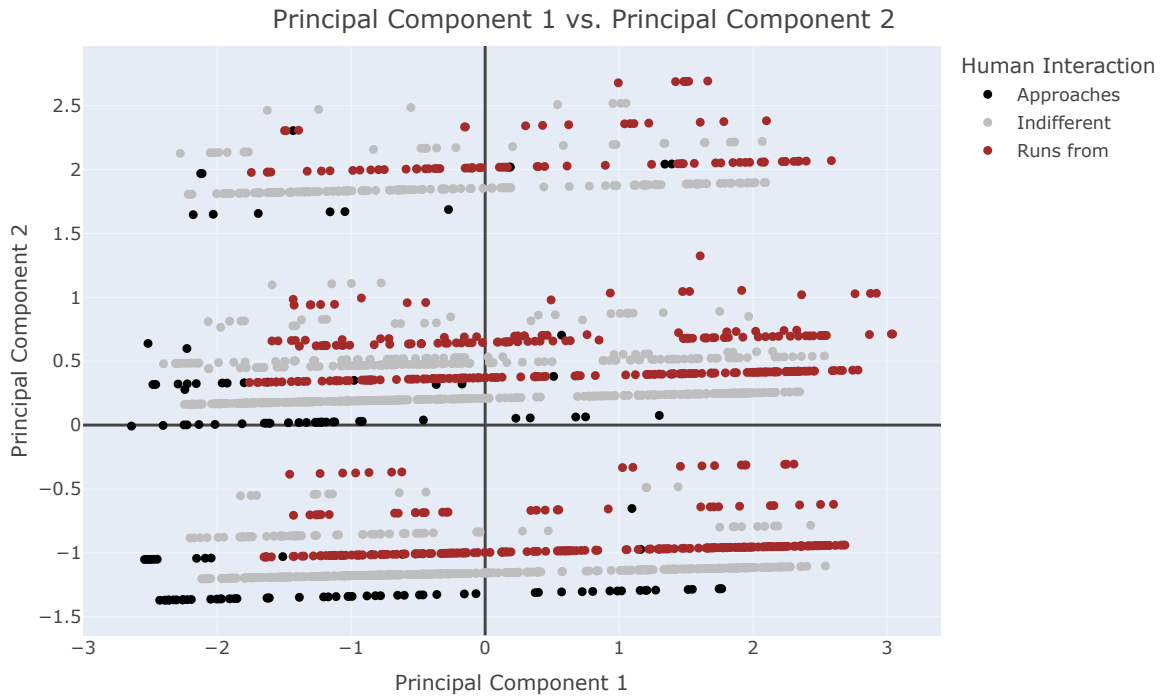


Figure 6 is a biplot of the PCA grouped by Human Interaction. `latitude` and `longitude` appear to be closely correlated, which is consistent with the previous findings. The near 90 degree angle by `shift_num`, `location_num`, and `latitude` and `longitude` vectors show that these variables are not closely correlated.

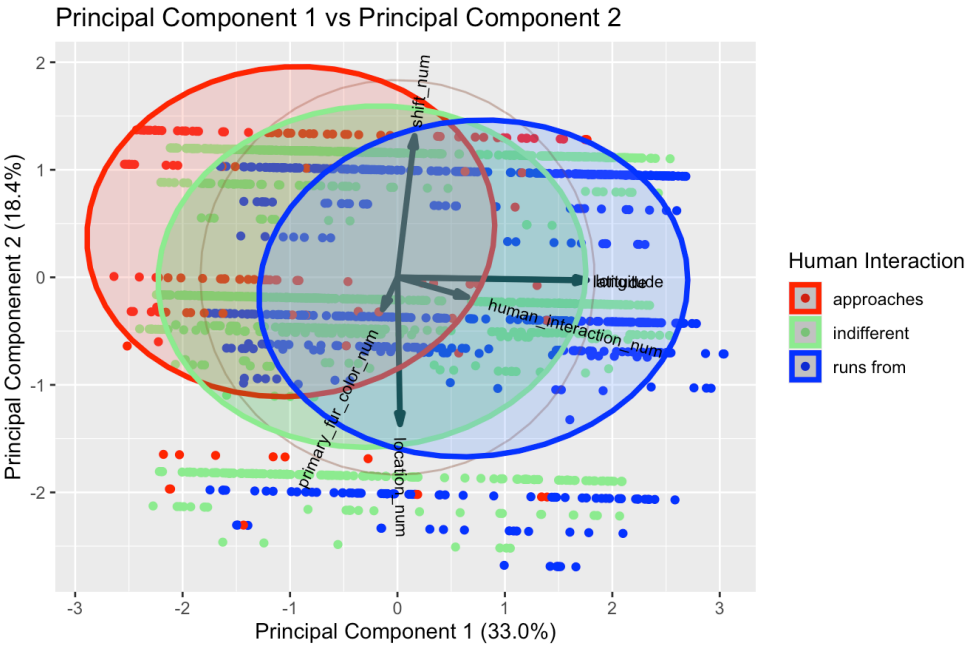
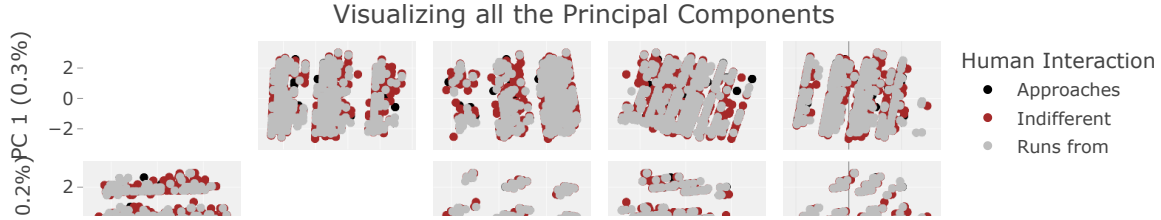


Figure 7 shows the result of comparing each principal component against each other. The points are colored based on Human Interaction. Upon first glance, principal component 3 and principal component 2 are clustered into very distinct, separate groups, potentially revealing a relationship between `primary_fur_color` (principal component 3) and `longitude` (principal component 2).



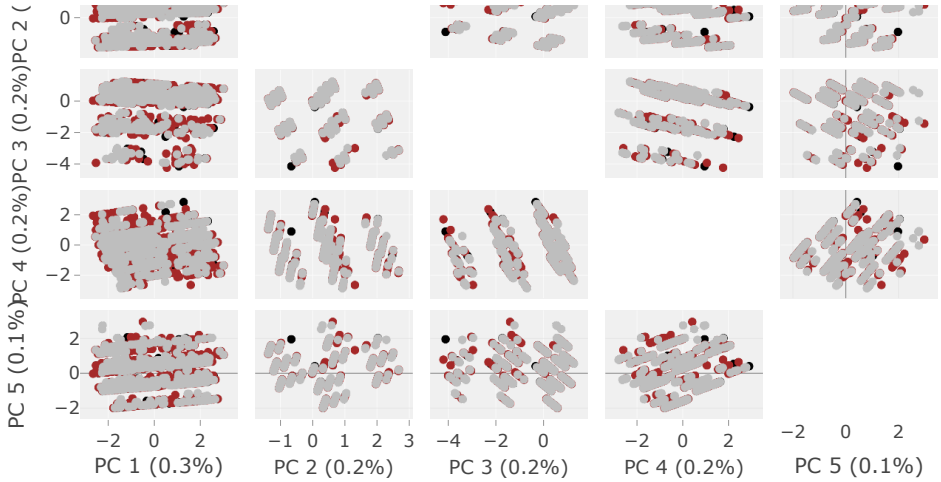
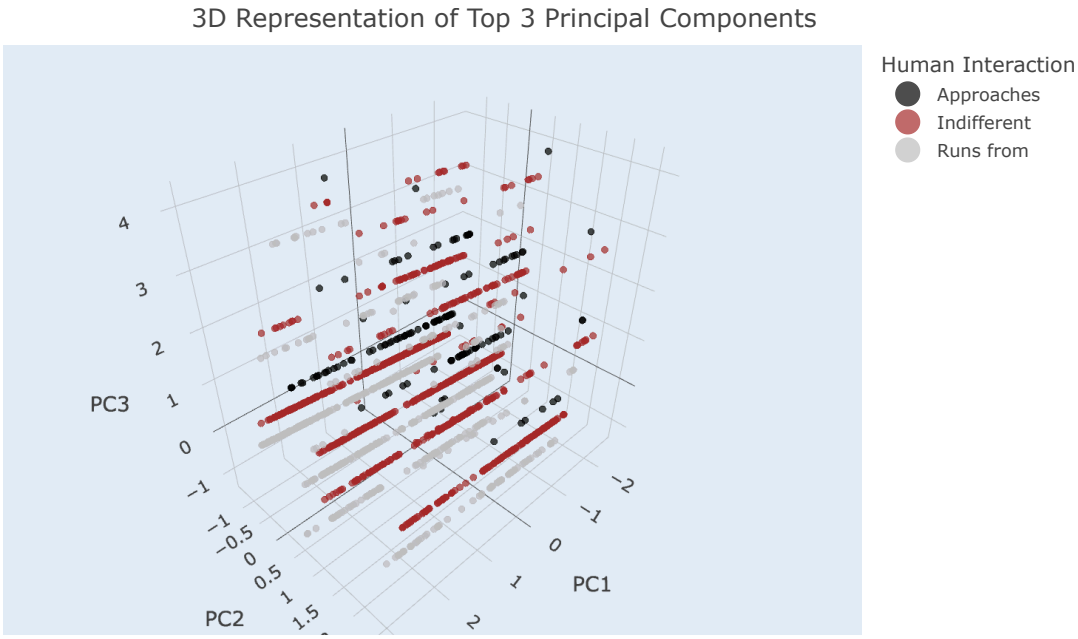


Figure 8 visualizes the top 3 principal components in 3D space. The points are colored based on Human Interaction. All the points run quite uniform in specific linear clusters when looking at the plot from different angles.



Chi-Squared Testing

After conducting a chi-squared test between `hectare` (Geographic Location) and `human_interaction`, `location` and `human_interaction`, `shift` and `human_interaction`, and `primary_fur_color` and `human_interaction`, it was found that all these variables had a p-value less than 0.05, the set significance level, indicating that there is a significant correlation between of the two variables and to reject the null hypothesis.

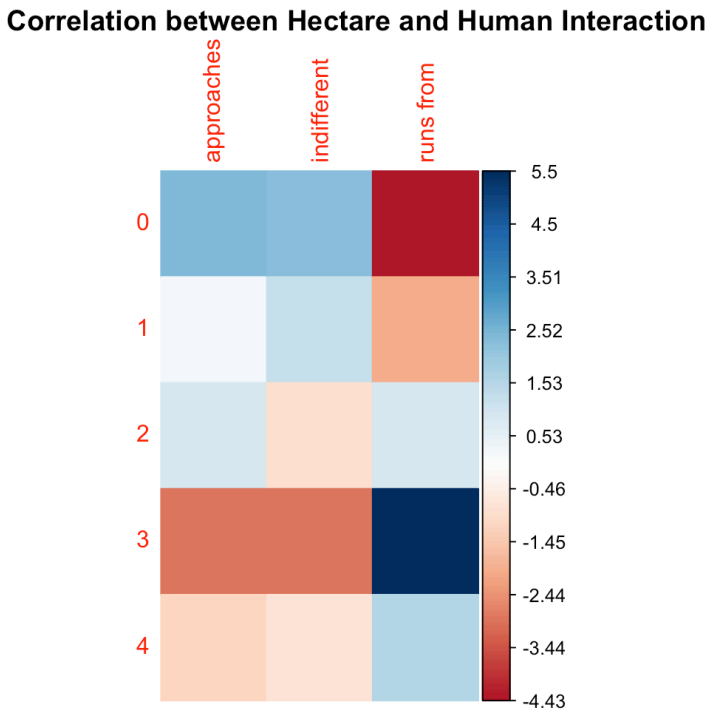
Hectare and Human Interaction

The p-value from the chi-squared test is

```
[1] 3.649006e-16
```

which is less than the set significance level of 0.05, meaning that there is a significant correlation between `hectare` and `human_interaction`.

Figure 9 below shows that squirrels in the 30A-39H hectare are more likely to run away from humans and that squirrels in the 01A-09I hectare are more likely to approach humans.



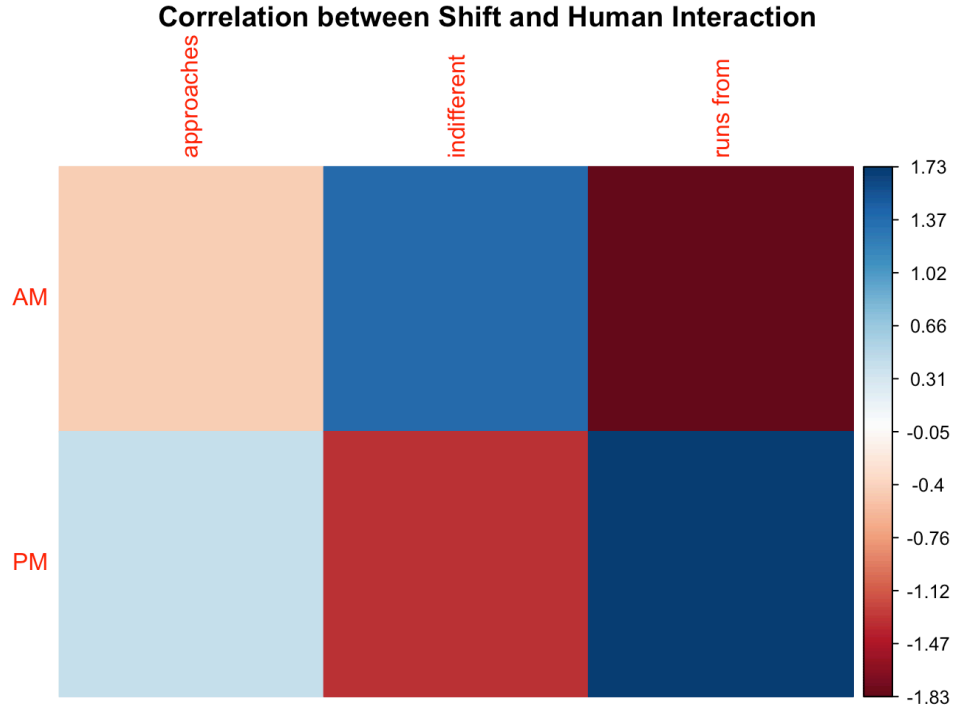
Shift and Human Interaction

The p-value from the chi-squared test is

```
[1] 0.005426958
```

which is less than 0.05, thereby indicating a significant correlation between `shift` and `human_interaction`.

Figure 10 below displays this correlation, showing that squirrels are more likely to run away from humans in the second half of the day (“PM”) and less likely to in the first half of the day (“AM”). Interestingly, most squirrels were found to be indifferent towards humans during the first half of the day (“AM”).



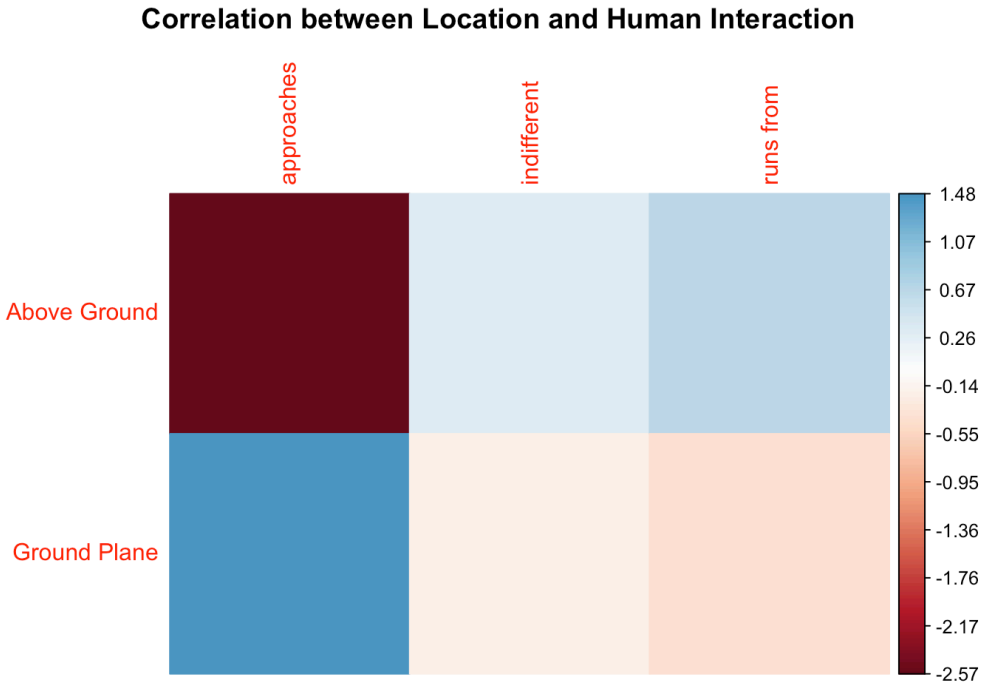
Location and Human Interaction

The p-value from the chi-squared test is

```
[1] 0.008341746
```

which is less than 0.05, thereby indicating a significant correlation between `location` and `human_interaction`.

Figure 11 shows this correlation between `location` and `human_interaction`. Squirrels tend to be less likely to approach humans when they are "Above Ground," understandable since they are not on the same plane as humans, the "Ground Plane."



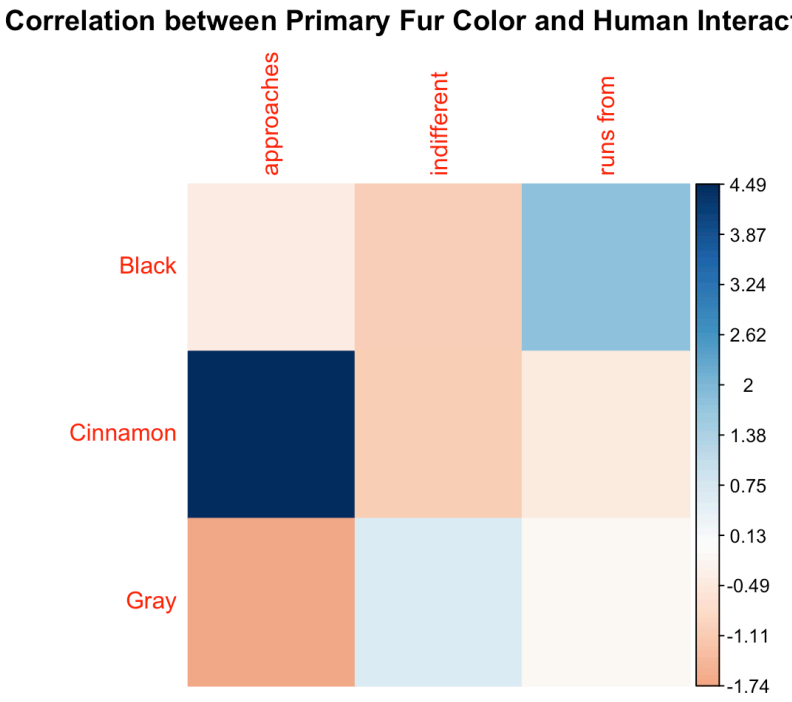
Primary Fur Color and Human Interaction

The p-value from the chi-squared test is

```
[1] 6.31357e-06
```

which is less than 0.05, meaning that there is a statistically significant association between `primary_fur_color` and `human_interaction`.

Figure 12 shows this correlation, with squirrels with a "Gray" fur color tending to have slight inclination to not approach humans and squirrels with a primary fur color of "Cinnamon" more likely to approach humans, surprising since "Cinnamon" squirrels make up a minority of the dataset.



K-Nearest Neighbors (KNN) Model

This K-Nearest Neighbors (KNN) Model will predict a squirrel's interaction with humans based on `hectare`, `location`, `shift`, and `primary_fur_color` (k = 5).

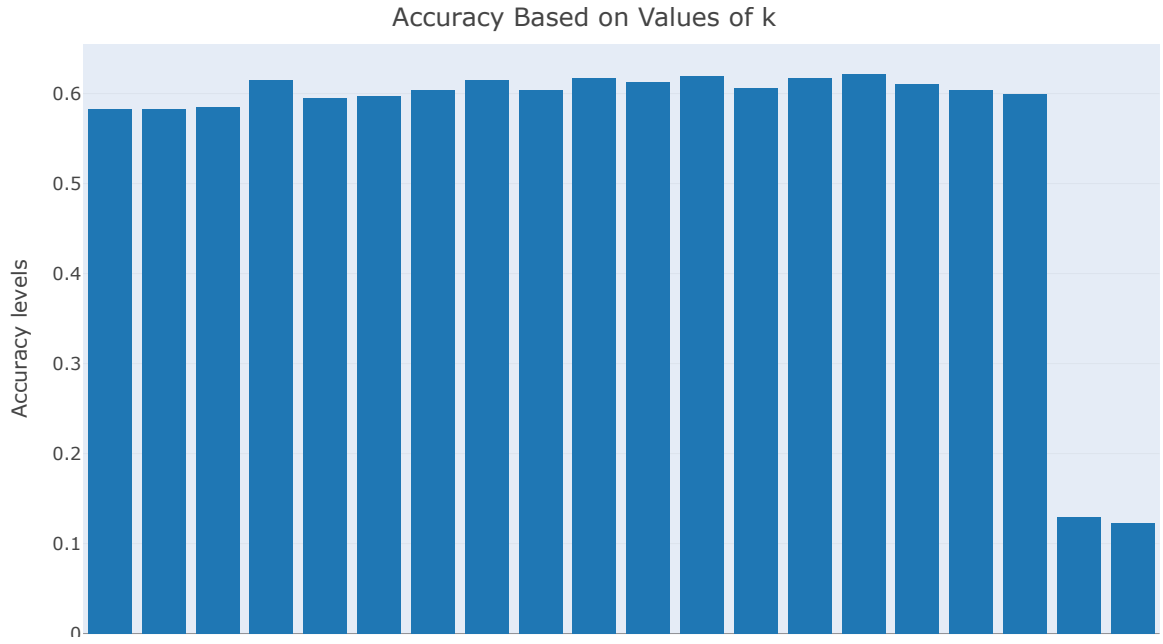
Here is the confusion matrix (Table 3)

```
             pred
              approaches indifferent runs from
  approaches           4          19         5
  indifferent          9         224        45
  runs from            1          99        33
```

and the accuracy.

```
[1] 0.594533
```

Figure 13 shows the accuracy of different KNN models run at different values of k to see which one yields the highest accuracy.



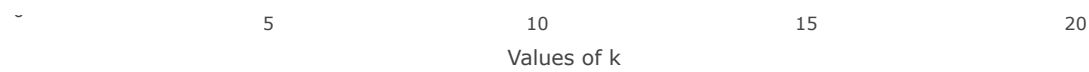


Figure 13 shows that the plot has the highest accuracy of 0.6218679 at k = 15.

Running the model again at k = 15, we can see that in the confusion matrix (Table 4), only the number of correct "indifferent" predictions increases, while the number of true accurate "approaches" and "runs from" predictions decreases, which could be a result of overfitting the model and the training and testing data including more observations in the "indifferent" categories.

```
             pred
              approaches indifferent runs from
  approaches           0          23         5
  indifferent          3         246        29
  runs from            0         106        27
```

The accuracy does increase, but that's due to the number of correct "indifferent" predictions increasing at the cost of correct predictions for "approaches" and "runs from" decreasing.

```
[1] 0.6218679
```

Discussion

As one of few undomesticated animals in New York City, the Eastern gray squirrel's adaptation is a fascinating phenomena of evolution (Allen, "Getting to Know"). However, only so much can be deduced about them just be looking at the Eastern gray squirrel's solely in Central Park, New York City. While it is now known Geographical Location, Location, Shift, and Primary Fur Color all have a significant correlation with Human Interaction, we can only generalize these results to the Eastern gray squirrels in Central Park. But does it hold true for them in New York City parks and for all Eastern gray squirrels in general? More data on the Eastern gray squirrels in other locales in New York City, as well as the original species of Eastern gray squirrels, will be needed to draw more conclusions about the evolutionary pathway of the species, a worthwhile goal to understand the psyche of the Eastern gray squirrel. Ramifications of this research will include better understanding of the Eastern gray squirrel, as well as knowledge to increase the safety of interactions between squirrels and humans so they can continue to coexist peacefully.

Code and Data Availability

To access the dataset the 2018 Central Park Squirrel Census, go to NYC Open Data website here: https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw/about_data

For more insight into my data cleaning and analysis process, go to my GitHub repository here: https://github.com/the-codingschool/DSRP-2024-Devyani/tree/main.

Acknowledgements

I would like to thank my mentor, Ms. Devyani Rastogi, for guiding me during the research process and providing the dataset. I would also like to thank Ms. Sarah Parker for teaching the lectures and providing me with all the skills and knowledge necessary to complete this project. Additionally, thank you to my teaching assistant Ms. Wanjiru Randolph and Ms. Pippa Lother for helping me troubleshoot through technical difficulties. Finally, I would like to thank The Coding School and Columbia University for hosting this program and making this research possible. You all have instrumental in my research and data science journey, and I could not thank you enough.

Sources

2020 Decennial Census. "New York City, New York." United States Census Bureau, 2020, data.census.gov/profile/New_York_city,_New_York?g=160XX00US3651000#populations-and-people. Accessed 15 Aug. 2024.

Allen, Jamie. "Getting to Know Central Park's Squirrels." Central Park Conservancy, Oct. 2022, www.centralparknyc.org/articles/getting-to-know-central-parks-squirrels#:~:text=Plugging%20the%20numbers%20into%20a,(1%2C798%20per%20square%20mile).&text=Based%20on%20previous%20counts%20we,be%20a%20healthy%2C%20sustainable%20number. Accessed 15 Aug. 2024.

—. "2018 Central Park Squirrel Census." The Squirrel Census, 2012, www.thesquirrelcensus.com/. Accessed 15 Aug. 2024.