Investigating Factors into Squirrels' Attitudes towards Humans in New York

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

2023-11-08

Introduction and Data

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.

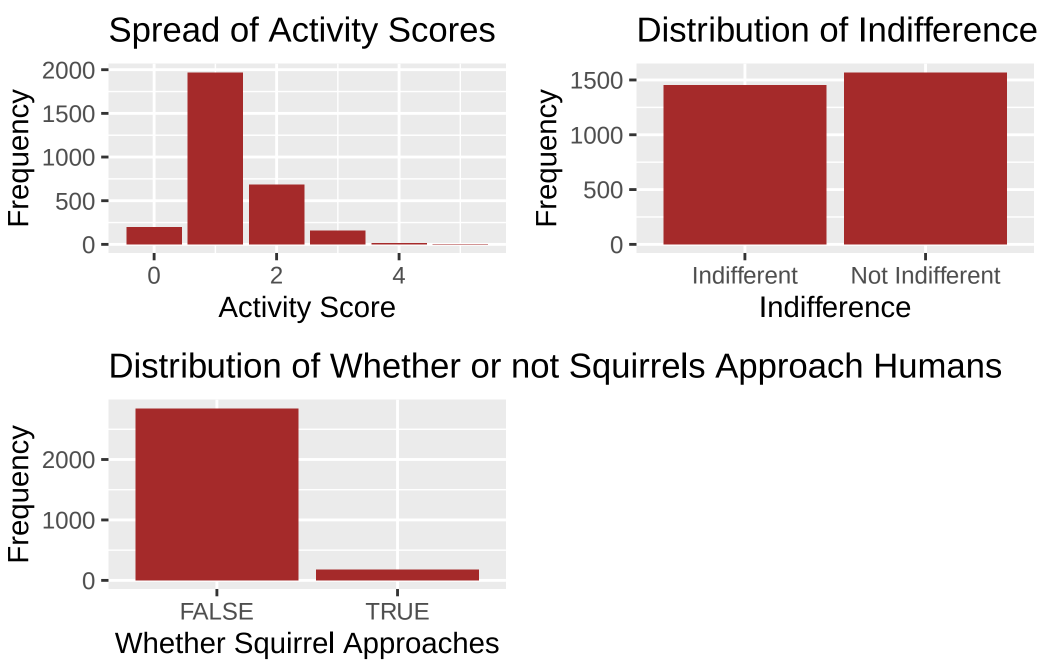
In October of 2018, the Squirrel Census Team and a group of over 300 volunteers collected the data based on squirrel sightings around Central Park.

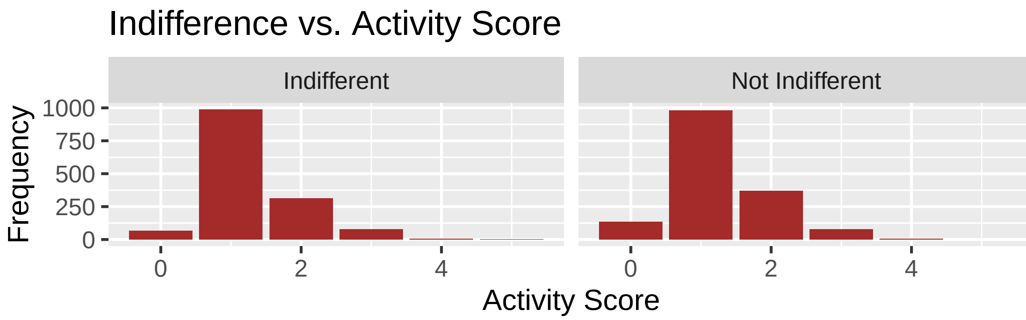
The dataset has 3023 observations and 31 variables, and it gives a wide range of observations and characteristics. It first gives us the location, in both longitude and latitude, the hectare of the park the squirrel was located in, the date, and whether it was found in the AM or PM. It also assigns each 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 squirrel's response when facing humans, being either that the squirrel approached, was indifferent, ran away, or any other action.

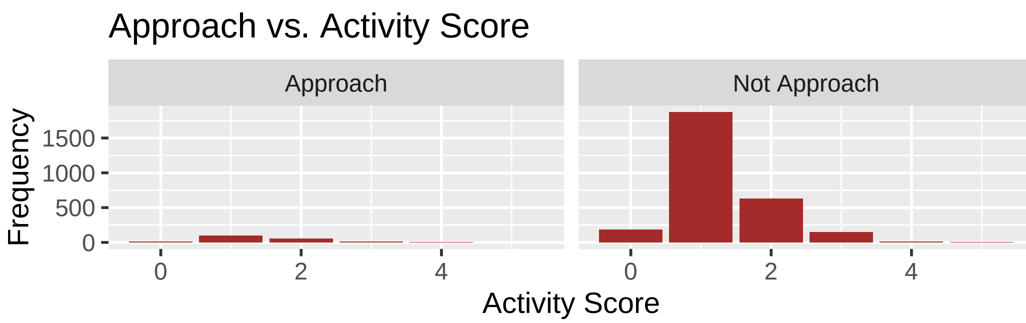
Initial exploratory data analysis

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.

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







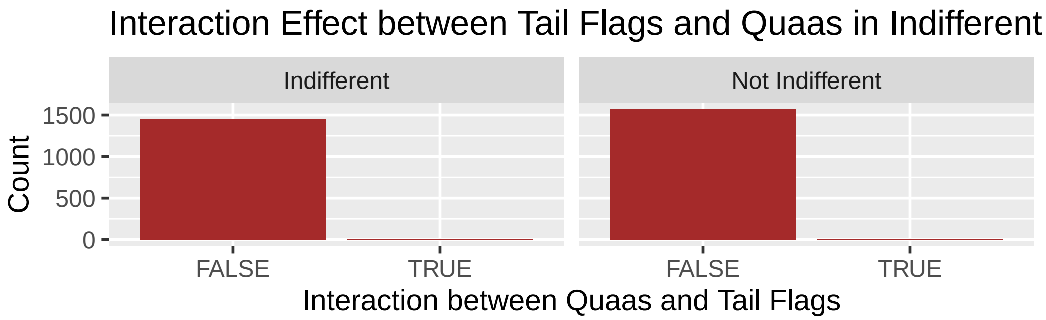
There seems to be roughly equal number of squirrels that were and weren't indifferent to the humans.

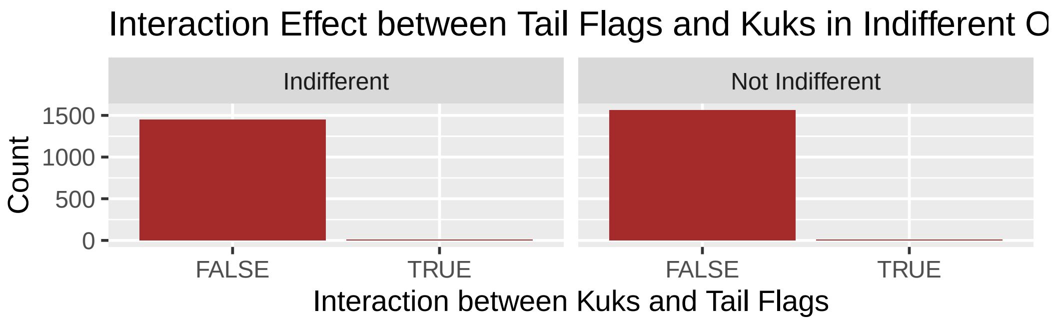
There were far fewer squirrels that approached humans than squirrels that didn't. This could show that they are not as synurbanized as we thought.

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 activities, specifically `tail flags`, when interacting with `Quaas`/`Kuks`, influences the likelihood that a squirrel is indifferent to humans. When squirrels flag their tails, the motion exaggerates their size and is used to confuse rivals or predators. When squirrels are heard quaaing, they are making an elongated vocal communication to indicate the presence of a ground predator. Meanwhile, kuk is a chirpy vocal communication used for a variety of reasons. Therefore, seeing the interaction between tail flags and quaas, more broadly, kuks, could help us understand whether or not the squirrels are indifferent to humans or not. In both graphs below, there are more counts of false interaction between both `Tail Flags` and `Quaas` as well as `Tail Flags` and `Kuks` (1600) than true interaction (1400), which shows a mild relationship between the variables.





Data dictionary

The data dictionary can be found here.

Methodology

Analysis approach

Our response variable is `Indifferent`, which is a categorical variable that indicates whether or not the squirrel is indifferent to humans. 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 logistic regression. We are using logistic regression because our dependent variable, `Indifferent,` is categorical. We will compare logistic regression models using 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 for model comparison.

Model 1

Model 1 includes all variables stated in our hypothesis.

Recipe for Model 1 steps: 1. Change response variable into factors. 2. Map all "FALSE" in `Above Ground Sighter Measurement` variable to 0, then convert the variable type to integer. 3. Create dummy variables for all nominal predictors. 4. Create interaction terms between `Quaas` and `Tail flags` 5. Remove all variables with zero variance.

```
Rows: 2,418
Columns: 14
$ Activity_Score                     <dbl> 1, 2, 2, 2, 1, 3, 1, 1, 1, 2, 2, 1,~
$ `Above Ground Sighter Measurement` <int> 0, 25, 0, 31, 15, 0, 0, 0, 0, 0, 0,~
$ Kuks                               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ Quaas                              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ Moans                              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ `Tail flags`                       <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, F~
$ `Tail twitches`                    <lgl> FALSE, FALSE, TRUE, TRUE, FALSE, TR~
$ Indifferent                        <fct> FALSE, FALSE, FALSE, TRUE, FALSE, F~
$ Age_Adult                          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 0, 1~
$ Age_Juvenile                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 1, 0~
$ `Primary Fur Color_Cinnamon`       <dbl> 1, 0, 0, 1, NA, 0, 0, 0, 0, 1, 1, 0~
$ `Primary Fur Color_Gray`           <dbl> 0, 0, 1, 0, NA, 1, 1, 1, 1, 0, 0, 1~
$ Location_Ground.Plane              <dbl> 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1,~
$ `QuaasTRUE_x_\`Tail flags\`TRUE`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
# A tibble: 2 x 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy binary     0.533    10  0.0103 Preprocessor1_Model1
2 roc_auc  binary     0.537    10  0.0106 Preprocessor1_Model1
```

```
# A tibble: 1 x 2
  mean_aic mean_bic
     <dbl>    <dbl>
1    2765.    2843.
```

Model 2

As the EDA has shown, Activity score does not seem to have a relationship with `Indifferent`, so we take it out for Model 2. In addition, there is a strong correlation between `Location` and `Above Ground Sighter Measurement`, so we omit `Location` because `Above Ground Sighter Measurement` is more granular.

Recipe for Model 2 steps: 1. Change response variable into factors 2. Map all "FALSE" in `Above Ground Sighter Measurement` variable to 0, and not "FALSE" to 1, then convert the variable type to factors. 3. Create dummy variables for all nominal predictors 4. Remove all variables with zero variance

```
Rows: 2,418
Columns: 12
$ `Above Ground Sighter Measurement` <int> 0, 25, 0, 31, 15, 0, 0, 0, 0, 0, 0,~
$ Kuks                               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ Quaas                              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ Moans                              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ `Tail flags`                       <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, F~
$ `Tail twitches`                    <lgl> FALSE, FALSE, TRUE, TRUE, FALSE, TR~
$ Indifferent                        <fct> FALSE, FALSE, FALSE, TRUE, FALSE, F~
$ Age_Adult                          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 0, 1~
$ Age_Juvenile                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 1, 0~
$ `Primary Fur Color_Cinnamon`       <dbl> 1, 0, 0, 1, NA, 0, 0, 0, 0, 1, 1, 0~
$ `Primary Fur Color_Gray`           <dbl> 0, 0, 1, 0, NA, 1, 1, 1, 1, 0, 0, 1~
$ `QuaasTRUE_x_\`Tail flags\`TRUE`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
# A tibble: 2 x 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy binary     0.524    10  0.0136 Preprocessor1_Model1
2 roc_auc  binary     0.539    10  0.0120 Preprocessor1_Model1
```

```
# A tibble: 1 x 2
  mean_aic mean_bic
     <dbl>    <dbl>
1    2763.    2829.
```

Model Conditions

We will check model conditions for Model 2 since this is the better performing model due to higher accuracy and AUC and lower AIC and BIC.

Linearity

There are no numeric variables in Model 2, so linearity is satisfied.

Randomness

The data was collected from the sightings of squirrels in Central Park, NYC from a group of volunteers. Although the volunteers are not randomly sampled, the sample of squirrels can be considered as random since we do not have reason to believe that the squirrels collected in this study differ systematically from squirrels in the rest of the world in regards to their indifference to humans. Therefore, randomness is satisfied.

Independence

The data are not spatially or time correlated since the data were collected in one time snapshot and in one location. Therefore, we have no reason to believe that independence was violated.

Thus, all model conditions for logistic regression is satisfied.

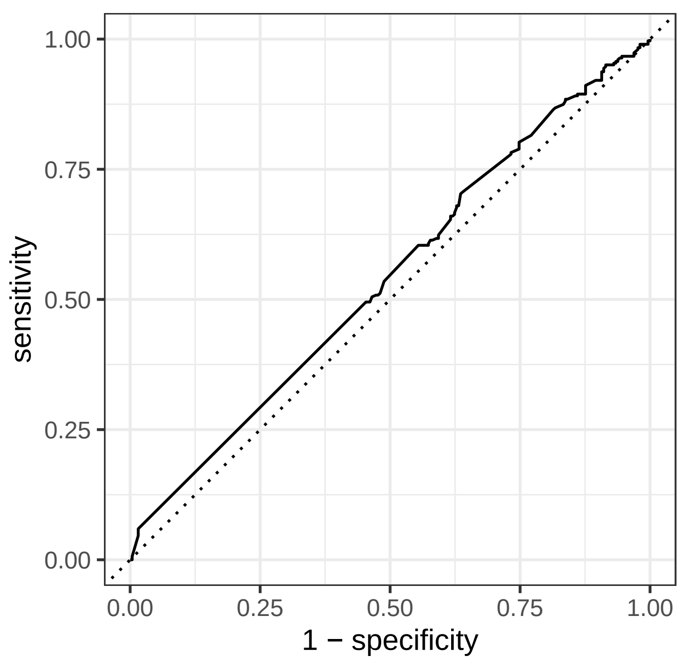
Results

Since Model 2 has a higher accuracy and AUC, as well as lower AIC and BIC, we will choose Model 2. Next, we fit the model to the entire squirrels dataset and interpret the coefficients.

```
# A tibble: 12 x 5
   term                                  estimate std.error statistic p.value
   <chr>                                    <dbl>     <dbl>     <dbl>   <dbl>
 1 "(Intercept)"                         -0.181     1.43      -0.127   0.899
 2 "`Above Ground Sighter Measurement`"  -0.00749   0.00399   -1.88    0.0600
 3 "KuksTRUE"                            -0.495     0.232     -2.13    0.0332
 4 "QuaasTRUE"                           -0.808     0.379     -2.13    0.0329
 5 "MoansTRUE"                          -11.4     197.        -0.0577  0.954
 6 "`Tail flags`TRUE"                     0.149     0.175      0.850   0.395
 7 "`Tail twitches`TRUE"                 -0.0394    0.108     -0.365   0.715
 8 "Age_Adult"                           -0.0342    1.42      -0.0241  0.981
 9 "Age_Juvenile"                        -0.250     1.42      -0.176   0.860
10 "`Primary Fur Color_Cinnamon`"         0.122     0.231      0.528   0.598
11 "`Primary Fur Color_Gray`"             0.298     0.210      1.42    0.156
12 "`QuaasTRUE_x_\\`Tail flags\\`TRUE`"   2.68      1.16       2.30    0.0214
```

```
# A tibble: 605 x 34
   .pred_FALSE .pred_TRUE     X     Y `Unique Squirrel ID` Hectare Shift    Date
         <dbl>      <dbl> <dbl> <dbl> <chr>                <chr>   <chr>   <dbl>
 1       NA         NA    -74.0  40.8 32E-PM-1017-14       32E     PM     1.02e7
```

```
 2      NA          NA      -74.0  40.8 11H-AM-1010-03       11H     AM     1.01e7
 3       0.533       0.467 -74.0  40.8 16I-AM-1008-01       16I     AM     1.01e7
 4       0.480       0.520 -74.0  40.8 22F-PM-1014-05       22F     PM     1.01e7
 5       0.536       0.464 -74.0  40.8 18A-PM-1018-01       18A     PM     1.02e7
 6       0.572       0.428 -74.0  40.8 17E-AM-1017-05       17E     AM     1.02e7
 7       0.480       0.520 -74.0  40.8 39C-PM-1006-01       39C     PM     1.01e7
 8       0.533       0.467 -74.0  40.8 6G-AM-1008-02        06G     AM     1.01e7
 9       0.533       0.467 -74.0  40.8 14F-AM-1007-05       14F     AM     1.01e7
10       0.480       0.520 -74.0  40.8 11B-PM-1014-05       11B     PM     1.01e7
# i 595 more rows
# i 26 more variables: `Hectare Squirrel Number` <dbl>, Age <chr>,
#   `Primary Fur Color` <chr>, `Highlight Fur Color` <chr>,
#   `Combination of Primary and Highlight Color` <chr>, `Color notes` <chr>,
#   Location <chr>, `Above Ground Sighter Measurement` <chr>,
#   `Specific Location` <chr>, Running <lgl>, Chasing <lgl>, Climbing <lgl>,
#   Eating <lgl>, Foraging <lgl>, `Other Activities` <chr>, Kuks <lgl>, ...
```




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.540
```

```
           Truth
Prediction FALSE TRUE
     FALSE    99  104
     TRUE    159  199
```

Of all the variables we examined, `KuksTRUE`, `QuaasTRUE`, `Above Ground Sighter Measurement`, and the interaction between `KuksTRUE` and `Tail flagsTRUE` were the only terms that had coefficients with significant p-values. The odds that a squirrel is indifferent to a human is multiplied by a factor of 1 if it kuks and 0 if it quaas. For every additional 1 meter in the squirrel's location above ground, the predicted log odds that it is indifferent decreases by `-0.1817`. In other words, kuks, quaas, and being above ground decrease the odds that a squirrel is indifferent to humans. However, the interaction term between `KuksTRUE` and `Tail flagsTRUE` has a positive coefficient, meaning the presence of both kuks and tail flags increases the predicted odds that a squirrel is indifferent: specifically, by a factor of 15. This is an interesting result because in the scientific literature, both kuks and tail flags are used by squirrels to warn other squirrels of a potential ground threat, indicating that the squirrels do see human presence as a threat, yet are indifferent.

The model does not have a high predictive power as shown by the low AUC (0.528) and accuracy ((121+173) / (121 + 130 + 137 + 173) = 0.5240642). This is somewhat expected given what we saw in the EDA section, where the distribution of the predictors is either similar across the two levels of response or very imbalanced.

Discussion and Conclusion

In this paper, we have investigated which factors significantly affect whether a squirrel is indifferent to human presence, using a logistic regression model and a 10-fold cross-validation for model selection. We concluded that based on our logistic model, there are 4 significant predictors of `indifferent`, which are `Kuks`, `Quaas`, `Above Ground Sighter Measurement`, and the interaction between `Kuks` and `Tail flags` since their p-values in the model is less than the $\alpha = 0.05$ significance level. Putting this into a larger context, this shows that the squirrels' attitude to humans are largely caused by human presence since according to the data descriptions, `Tail flags` are actions when squirrels see a threat, and it being true will increase the squirrels' odds of being indifferent. On the other hand, other variables are not related to human activities, and the presence of them will decrease the squirrels' odds of being indifferent. This shows that squirrels are more likely to be indifferent to humans if they see humans a threat and less like to be indifferent if they do not exhibit signs for threats.

One important limitations is the concerns about the way data is collected. For example, based on the data description, the data is collected from volunteer sighting in NYC. This might introduce some lack of rigorousness, especially when it comes to numeric variables like `Above Ground Sighter Measurement` since the volunteers might give a very subjective estimation of the height above the ground when standing and observing from distance. In addition, there are a lot of NA's in the dataset, which to some extent limits our ability to do analysis.

For future work, we will try to extract more useful predictors from feature selection and feature engineering steps in order to increase the predictive power of our model. For example, instead of simply dropping the longitude and latitude from the dataset, we could potentially search the longitude and latitude of downtown NYC and use this information to calculate the Euclidean distance (or perhaps Manhattan distance since it's in NYC) between each observation and the city center for the model. In addition, we could try some non-linear machine learning models, such as random forest and boosting, since the linear model seem not to be able to give accurate predictions in this case.

> **!** Important
>
> Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.