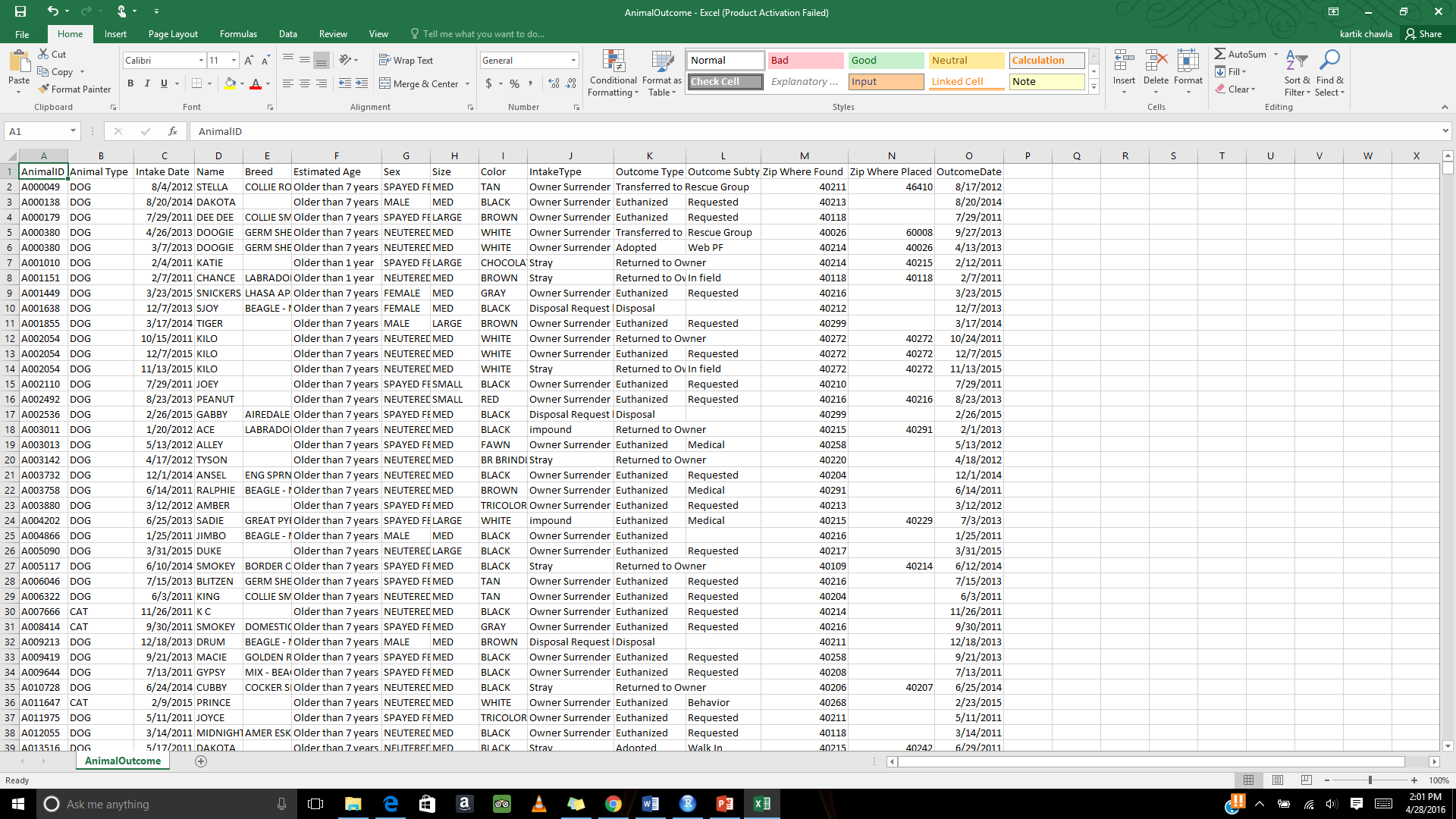
**The Dataset:**

We have selected a dataset that contains information on all the animals that are handled by animal shelters in Louisville, KY. The Dataset provides data on the types of animals, their breed, name, color, and sex, when they enter and leave the shelters, the reasons they enter and leave the shelters, and where they are found or taken in and where they are retrieved.

The dataset is located at <http://portal.louisvilleky.gov/latest_datasets?order=title&sort=asc>.

The data is structured as a CSV (comma separated variable) file with 15 attributes that are described below. There are 34,484 records in the file



**Descriptive Analytics**

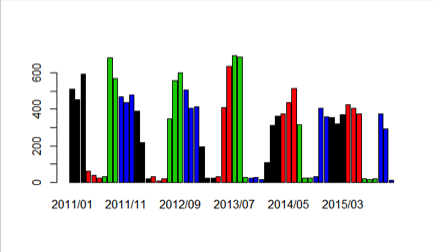
1. AnimalID: A unique code for each animal. This is a 6 digit number preceded by an A. No Null
2. AnimalType: The species of animal. String with options: Dogs, Cats, Birds, Guinea pigs, Livestock, Rabbit and others. No Null
3. IntakeDate: The date the animal enters the shelter. Multiple date formats 2011-2015. No Null
4. Name: The name of the animal. String ~1/3 Null
5. Breed: The breed of the animal. String ~80% Null
6. EstimatedAge: The animals estimated age. String with options Infant < 6months, Youth < 1year, >1year, >7years. ~30% Null
7. SimAge: This column was created to provide a more useful analytic. The age was randomly generated within the range provided by the EstimatedAge. Integer ~30% Null
8. Sex: The gender of the animal and if it is spayed or neutered. String with options Male, Female, Spayed, Neutered, or Unknown. 300 Null
9. Size: The animal’s size. String with options small, medium, large, xlarge, toy, puppy, or kitten. 878 Null
10. Color: The color of the animal. String 7 Null
11. IntakeType: The reason the animal was brought into the shelter. String with options Disposal Request, For Transport, Foster Return, Impound, Out Patient, Owner Surrender, Reported Found, Reported Lost, Stray. No Null
12. OutcomeType: How the animal left the shelter. String 98 Null
13. OutcomeSubtype: The reason for the OutcomeType. String ~1/3 Null
14. ZipWhereFound: The zip code for where animal was found/brought in. Integer Mostly between 40000-49999 863 Null
15. ZipWherePlaced: The zip code for where animal was placed. Integer ~50% Null
16. Outcome Date: The date of the animal left the shelter. Multiple date formats 2011-2015. No Null

For calculating the mean and standard deviation we need to have numerical values in the columns. The only attribute that contains numerical data for which mean and standard deviation can be calculated is the SimAge of the animal. And the mean age for all the animals in the dataset is and the Standard Deviation for age is

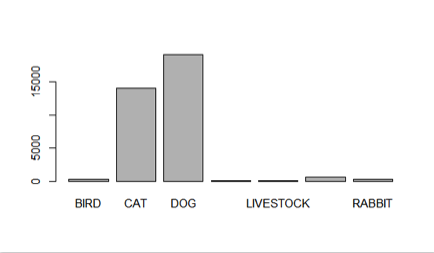
For many of the fields there are a relatively small number of missing values that can be omitted. For Age we plan to predict the age based on other attributes. For ZipWherePlaced, we can ignore missing data for OutcomeTypes Destroyed and Euthanized. This leaves ~10% missing which can be predicted based on outcome and intake zip. For attributes such as Breed, Name, and OutcomeSubtype, the attribute will not be needed for calculations.

**The Graphical Data:**

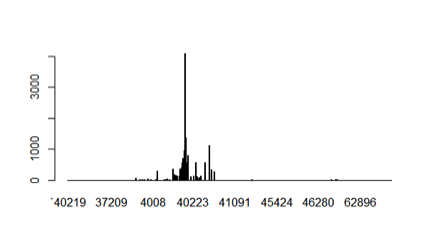
1. Graph for Strays per month for years from 2011 to 2015 colored by quarter



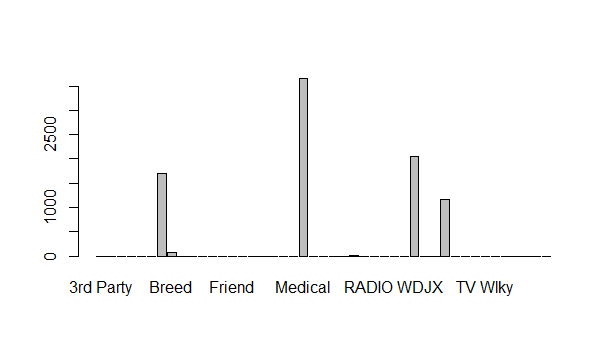
1. Graph for Adopted Animals by Animal Type



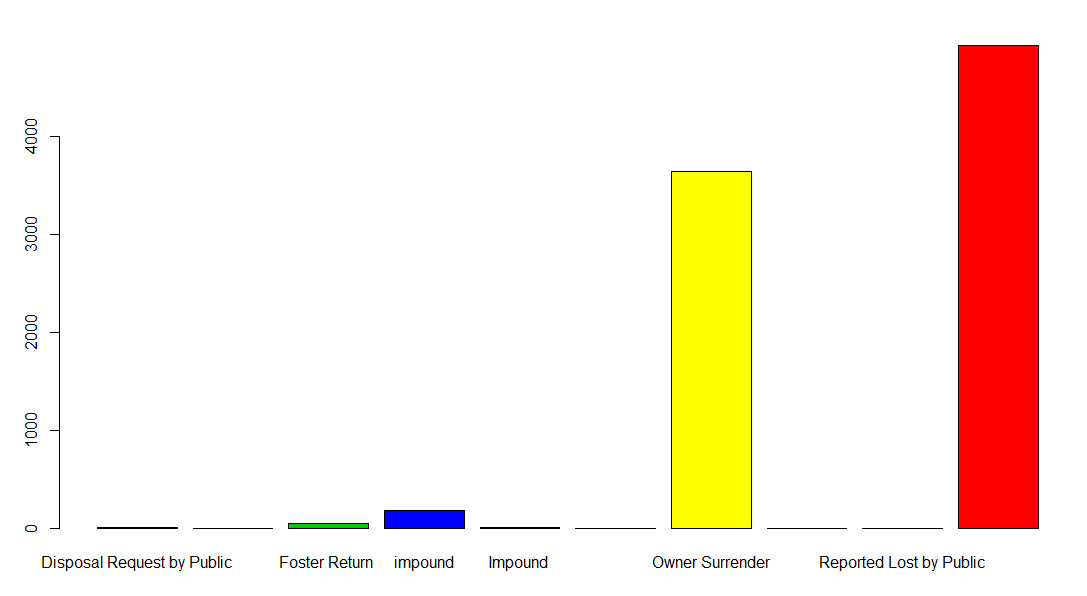
1. Graph for Stray Animals in every zip code



1. Euthanization of animals based on various reasons

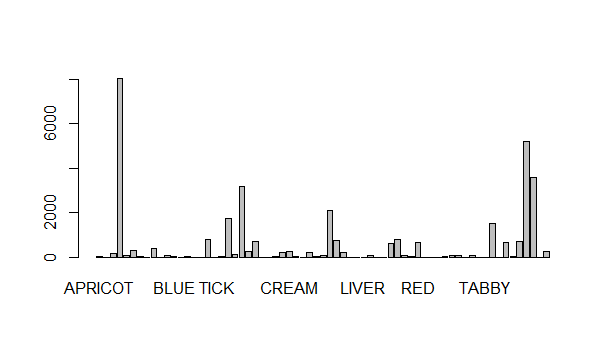


1. Euthanization based on Intake type



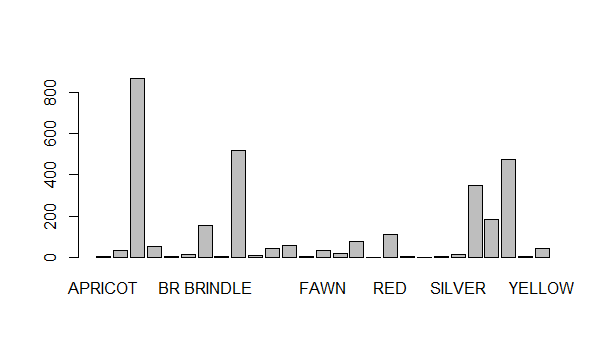
1. Which colored animal is adopted the most?

Color: Black- 8022



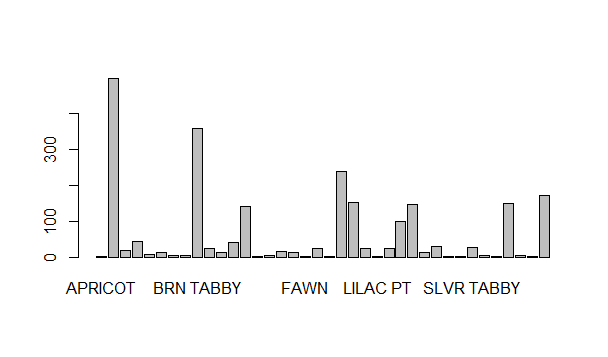
1. Which colored cat is most adopted?

Dog: Color- Black : 865



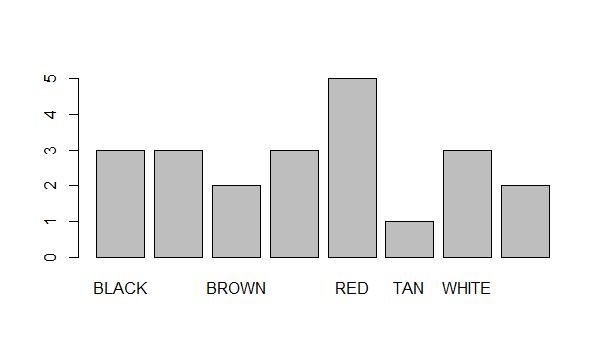
1. Which colored cat is adopted the most?

Color: Black- 496



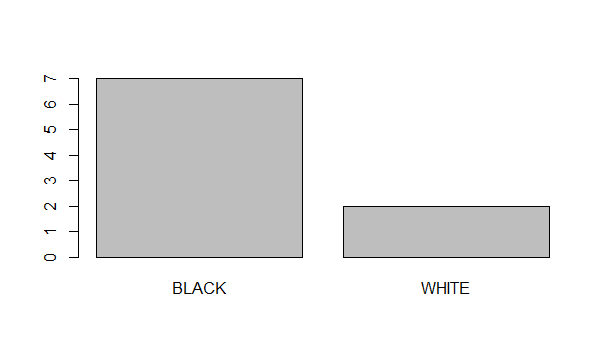
1. Which colored Bird is adopted the most?

Color: Red- 5

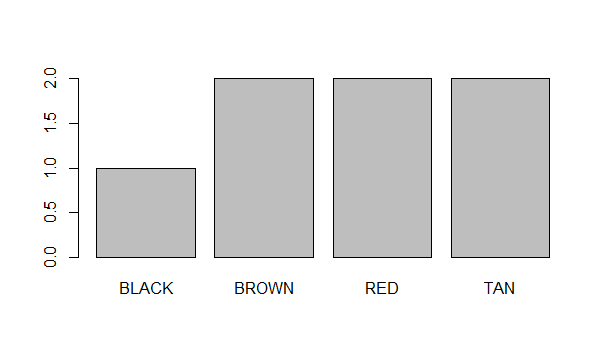


1. Similarly, for Livestock, Guinea pig and rabbit

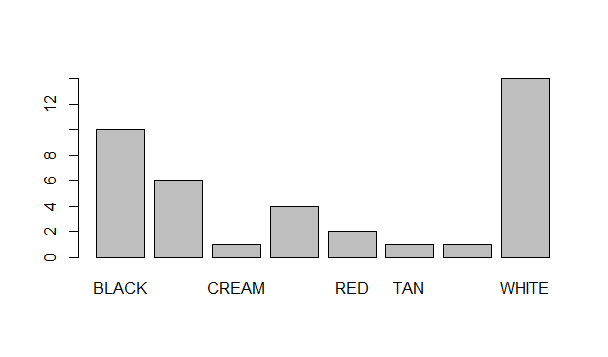
Livestock



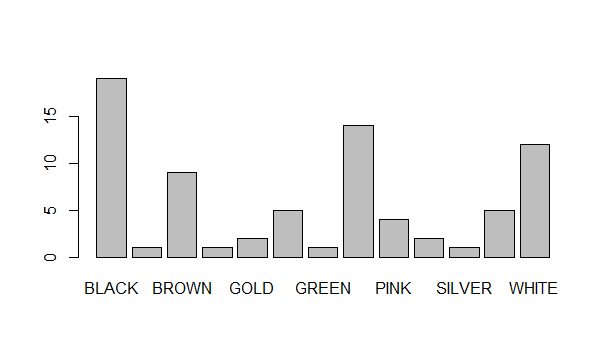
Pig:



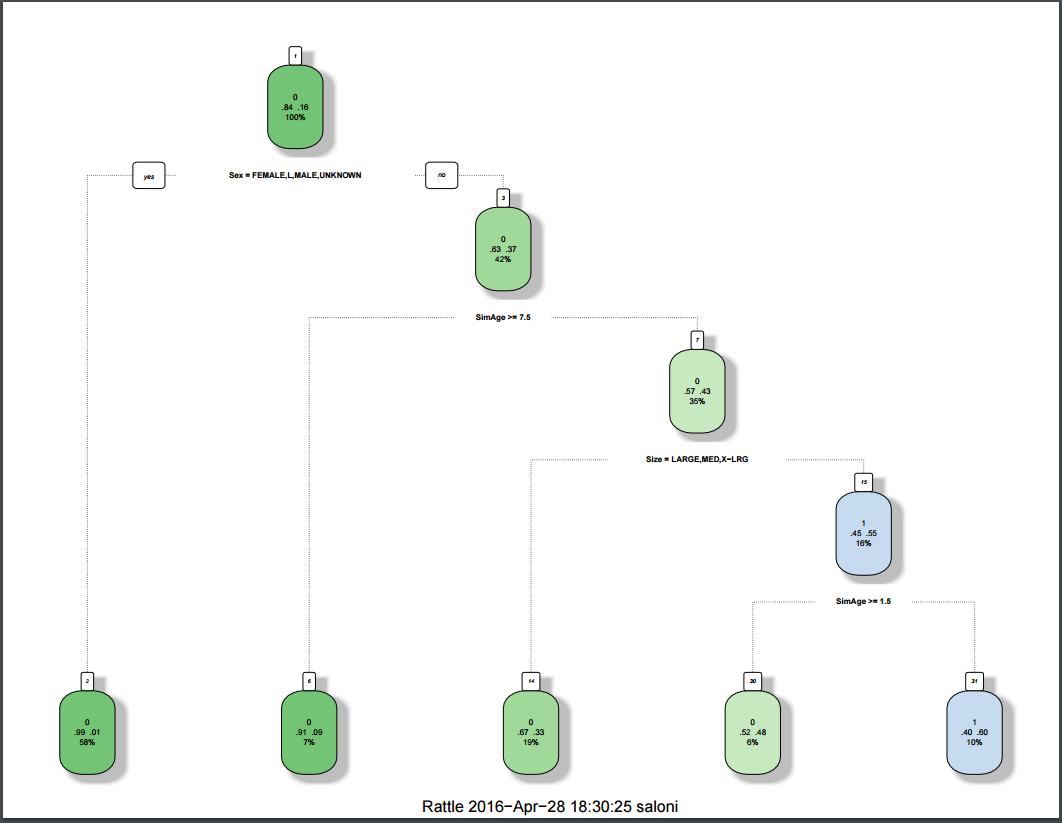
Rabbit:

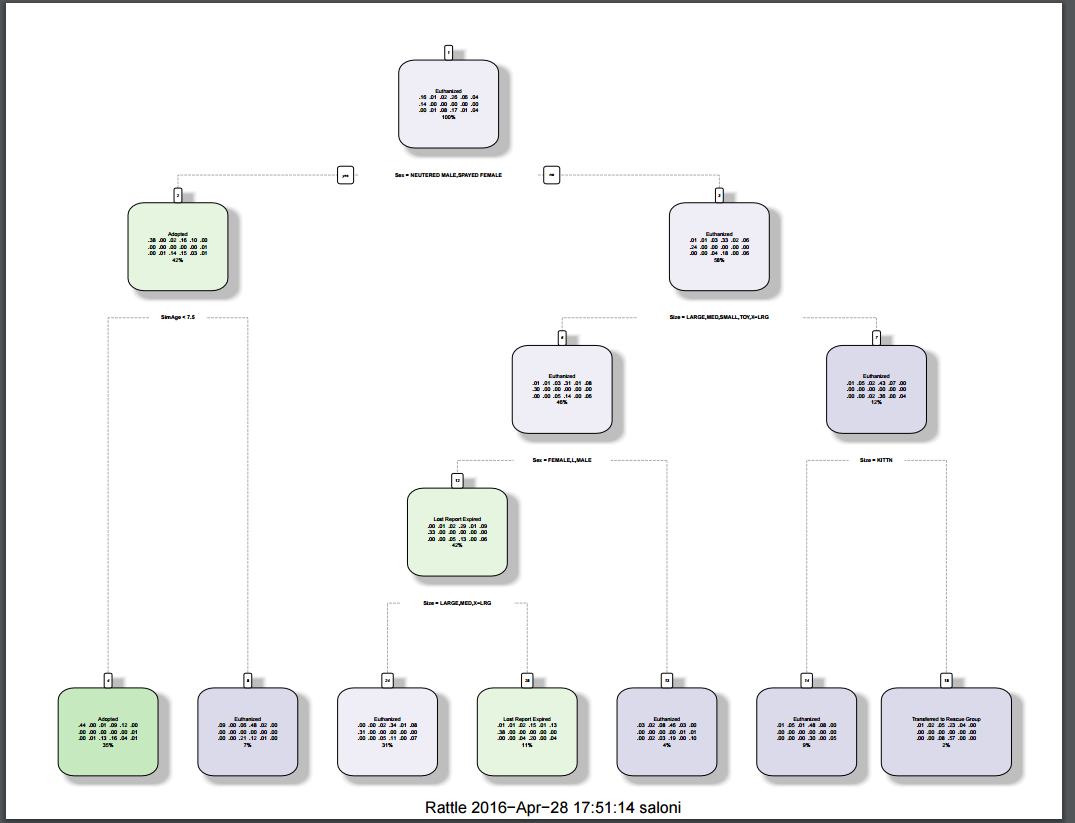


Others:

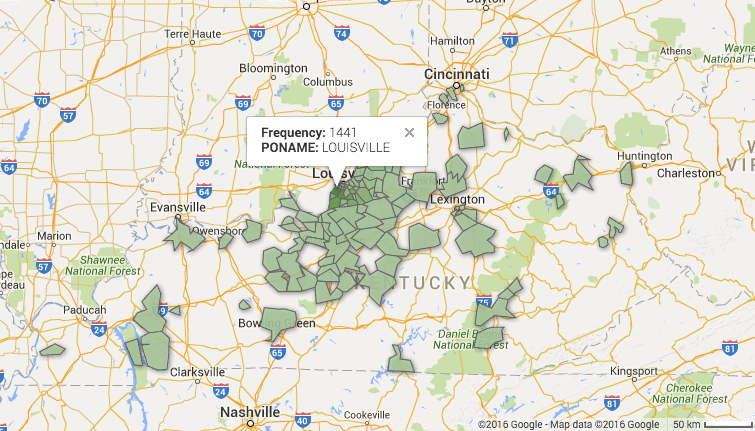


1. Adoption based on Size. Sex and age:





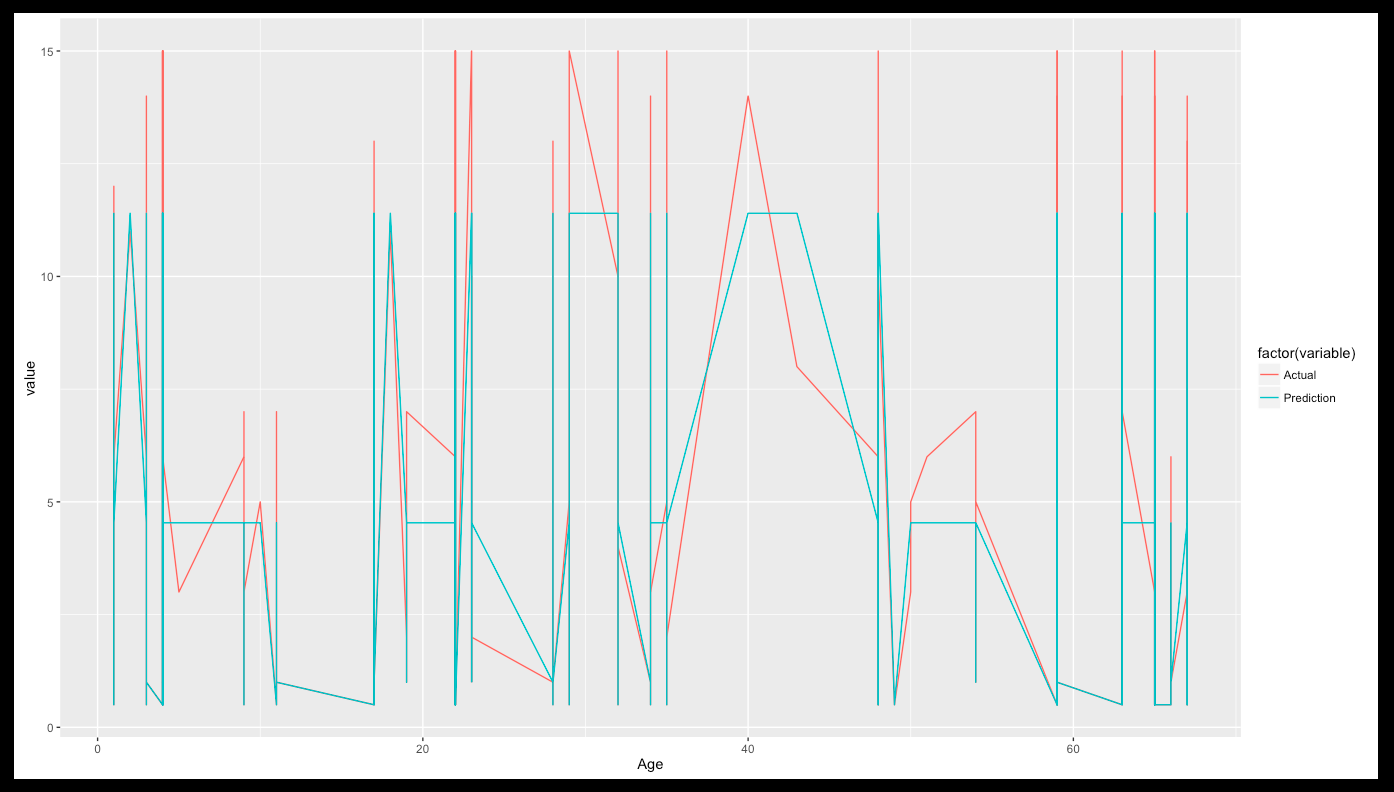
**Stray dogs by zip:**

****

**Predictive analysis:**

* What factors are most useful in determining whether an animal will be adopted?
* Determine which locations mostly have strays?
* Predict the colour of animal and types of animal being most adopted
* Predict age based on various attributes.

**Age Prediction:**



**Challenges faced:**

* The dataset, had unclear age of animals (column name: estimated age). On the basis of estimated age, we calculated the age of all the animals which were useful for us for all the analysis and prediction.
* The zip code attributes have some blank data and some mistyped values. We have cleaned the data by removing incorrect values or correcting the data. Once incorrect data was removed, we ignored blank records.
* The dates had various formats. We used Excel to format all the dates in the same way. This way we did set the format of dates as required for R functions.

**Problems**

Our goal is to determine what factors are most useful in determining whether an animal will be adopted. This will allow shelters to predict which animals will will be adopted. Also, if we are able to determine which locations are mostly to have adoptions of each type of animal, then shelters can transfer pets to where they are most likely to be adopted. We also want to predict age based on various attributes. This will allow shelters to better estimate the age of pets that are brought in.

**Challenges expecting to face during project**

1. The dataset, had unclear age of animals (column name: estimated age). On the basis of estimated age, we calculated the age of all the animals which were useful for us for all the analysis and prediction.
2. The zip code attributes have some blank data and some mistyped values. We will have to clean the data by removing incorrect data or correcting the data. Once incorrect data is removed, we can ignore blank data.
3. The dates have various formats. We can use Excel to format all the dates in the same way. This way we can set the format to the format required by R functions.

**Roles:**

Requirement gathering (Data set): Prateek Dua, Preston Rollins <- Searching the data set

Analysis: Chinmaya Khismatrao, Saloni Kankaria, <- Predicting and analysing the challenges in data set

Designing and coding: Preston Rollins, Prateek Dua

Testing: Saloni Kankaria, Chinmaya Khismatrao, Kartik Chawla

Documentation: Kartik Chawla, Preston Rollins