Seattle Pets Exploratory Data Analysis

Rebecca Hadi

3/30/2018

Background

While browsing for some data to play with, I stumbled across a data set that contained pet licenses for the city of Seattle. The data set can be found at https://www.kaggle.com/aaronschlegel/seattle-pet-licenses/data. Since I own (or rather, have a mutually beneficial relationship with) a 14-year old tuxedo cat named Jinx, I thought it would be fun to explore this data set.

Caveat: With any data found on the web, it is important to consider its method of collection and if there are any limitations/biases in the data set. Here, we are working with data from Kaggle, which says it's been gathered as part of the city's "Open Data Initiative". Any conclusions/observations from this analysis are specific to this data set and should not be extrapolated. The exploratory data analysis will likely shed more light on any limitations with the data set, but it's possible that there is bias within: (1) what data are included and/or (2) what proportion of pets in Seattle *are not* licensed and would therefore not be included in the data set (3) other things I'm not considering. Herein, any time I make a reference to "in Seattle", what I really mean is "in this Seattle data set which may or may not be representative". With that being said, we are only as good as what we can measure, and I'm sure there are learnings from this data set!

How much data are we working with?

When approaching a data set, I find that it's valuable to actually inspect the ranges of the data of which I'm working. While this is fairly intuitive, it is surprising that it's easy to miss something that could impact your analysis that would have been solved by a simple inspection of data data. An example of this that I frequently encounter professionally is dealing with NULL values, meaning, you get a data set, see the field you need, only to discover it's only populated for 30% of records.

In R, when working with a new data set, I run three statements right off the bat (assuming df is my data frame. * head (df) * summary (df) * str (df)

I won't show the results in this output, but this led me to observe that I have data going back to 2005. Nice! If I didn't go any further, my natural assumption would be that I have complete data over that time period. Let's see if that's true.

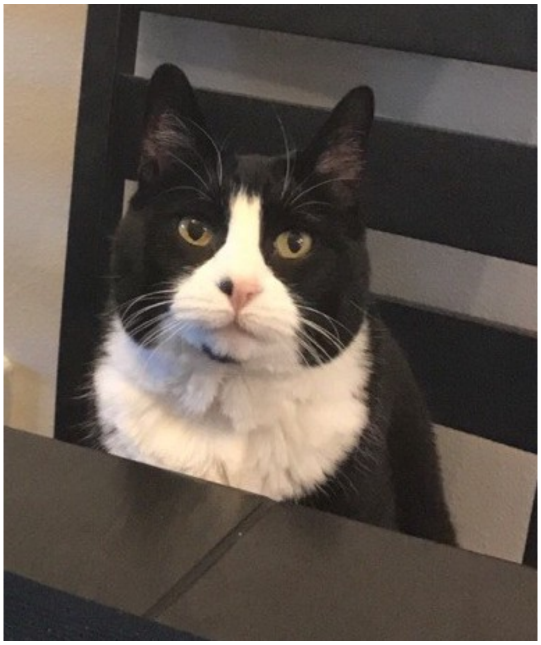
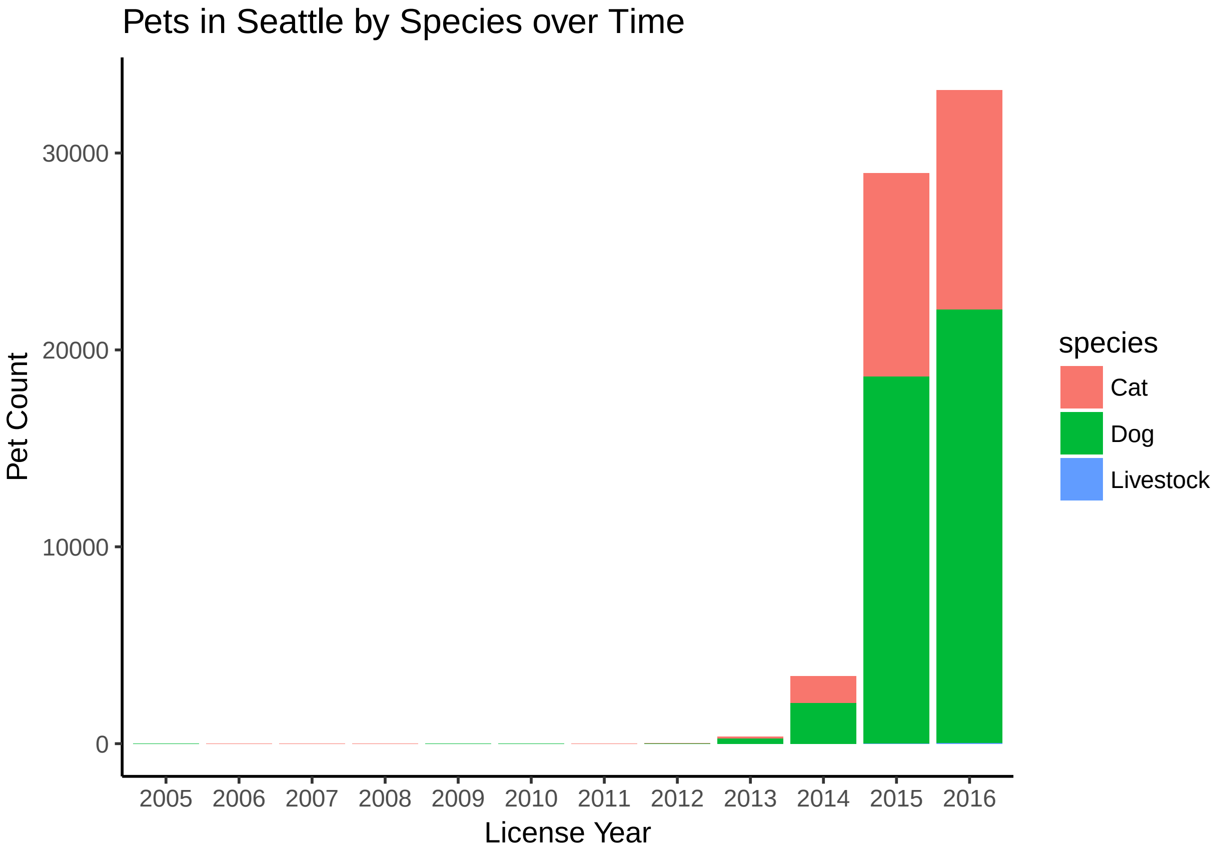


Figure 1: This is my cat, Jinx.



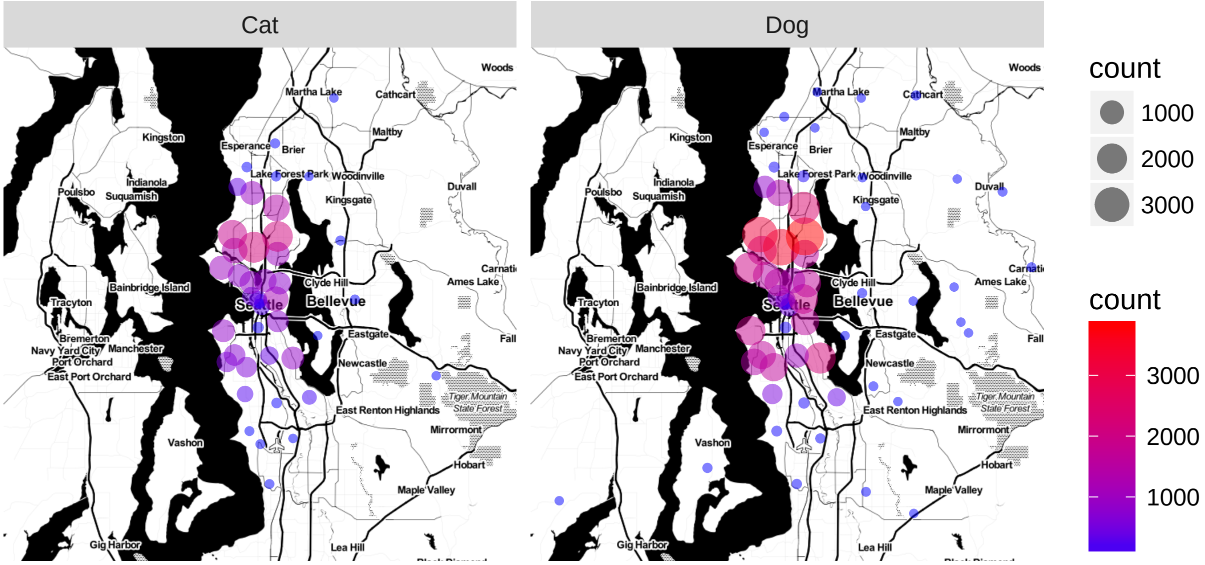
In the chart above, we see that the data from 2005 to 2012 is almost non-existent. It slowly picks up over 2013-2014, then a massive spike in 2015. It seems like there is some bias with the data collection, as it is unlikely there were minimal in Seattle until 2015. It's possible that only a minimal number of pets were *licensed* until 2015, but that also seems unlikely. I suppose we could do a bit of research to find out when Seattle started licensing or enforcing licensing of pets, but for the purpose of this analysis, I'm going to consider that this data set captures recent licensing behavior. It's difficult for me to evaluate if there were truly more pets licensed in 2016 vs. 2015, or if the data collection was better in 2016. I think the conclusion I'm going to work with is that this data set is not well suited to analyze trends. Nothing wrong with accepting limitations of your data set!

From this view, we can also see that the majority of the pets licensed in Seattle are dogs. Overall, 65 % of pets in the data set are dogs. The table below shows the counts and proportions for the data set.

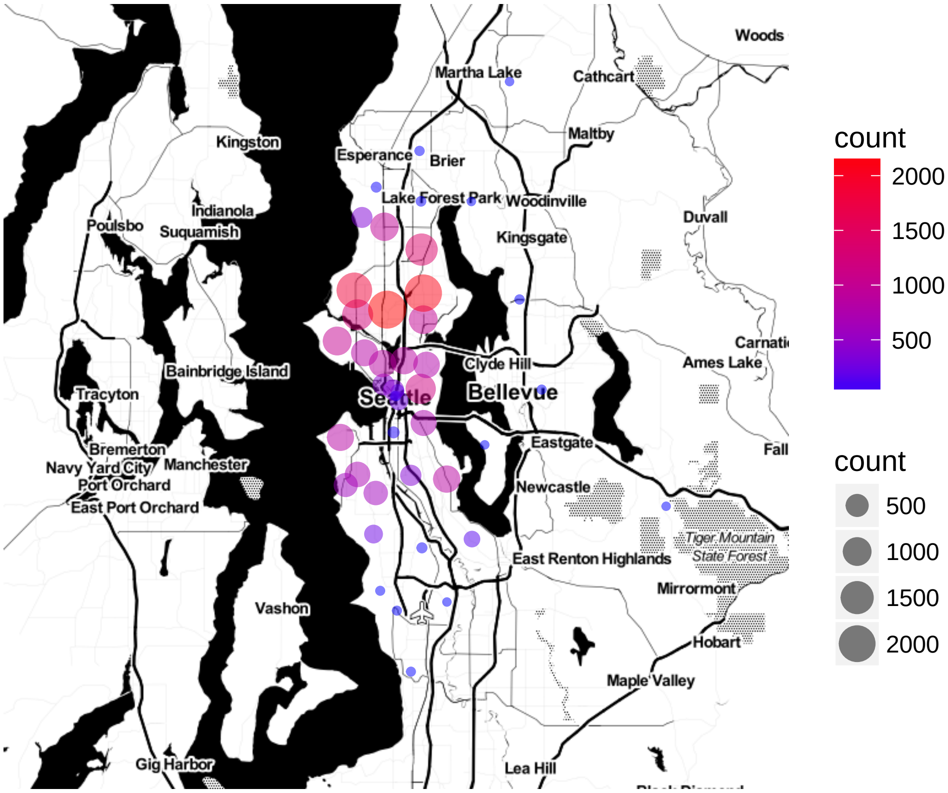
species	count	pct_total
Cat	22915	35
Dog	43076	65
Livestock	51	0

Spatial Analysis

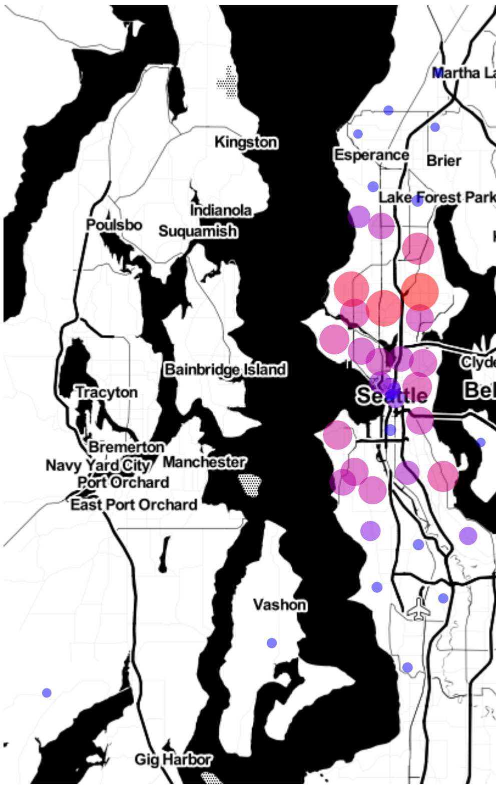
Total Pets in Seattle



Cats in Seattle



Dogs in Seattle



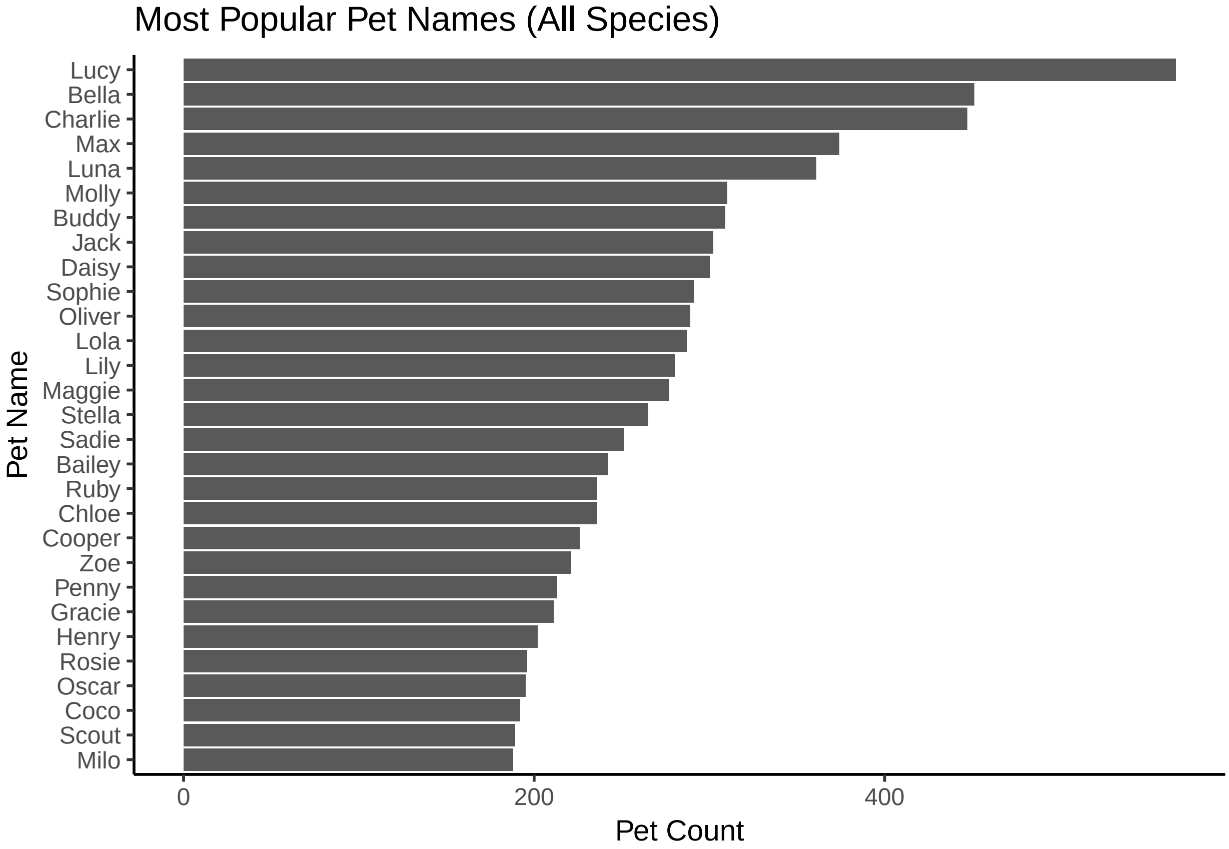
After some minor cleanup of the zip code data, I'm able to match 99% of records to a zip code in Seattle. Upon inspection, some zip codes were not in Seattle (presumably an error) or were not 5 digits long. Some zip codes were initially 9 digits (5 digit zip + 4 digit suffix). I was able to use a sub-string function to remove the last 4 digits to enable the join against the zip package.

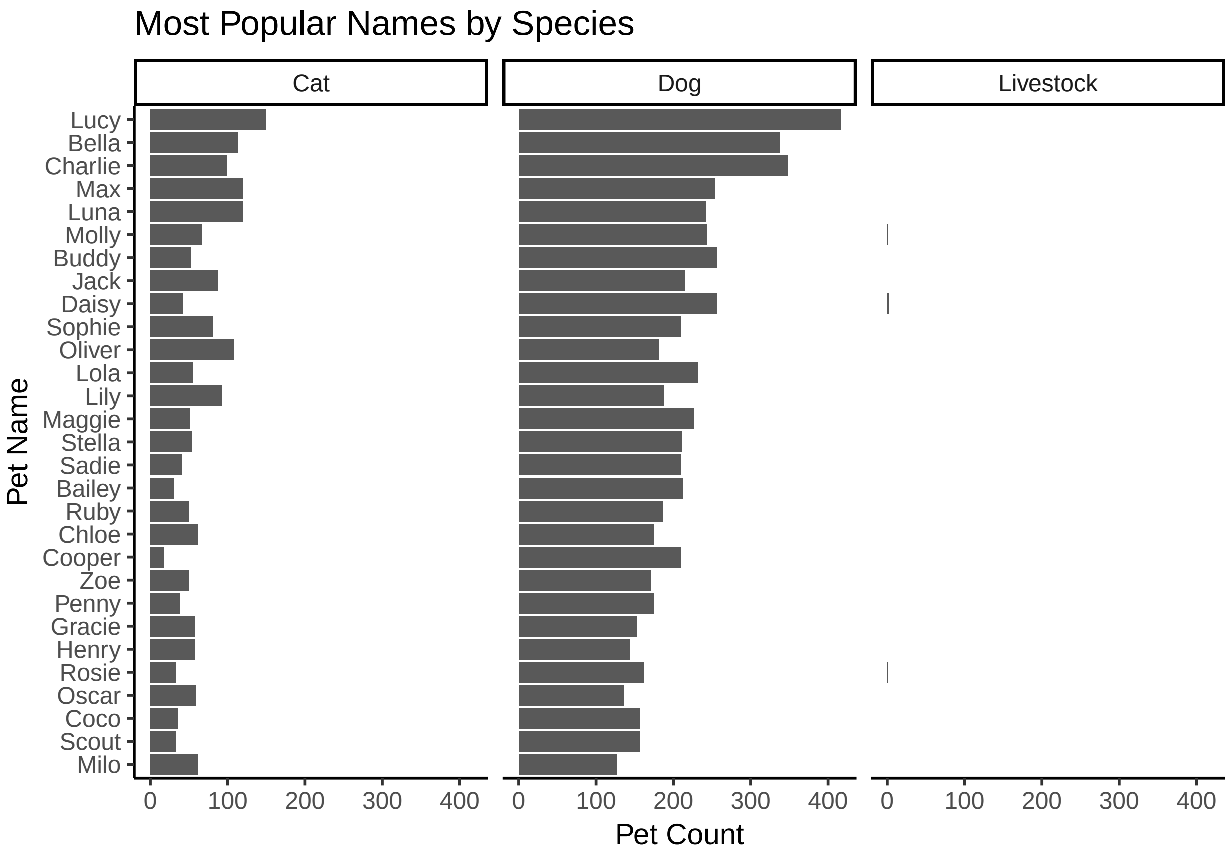
Based on these maps, we observe that there is not much of a difference in the concentration of cats compared to dogs. With this data set, it is difficult to discern if there are truly more dogs than cats or if there is some inherit bias with the pets that end up being licensed.

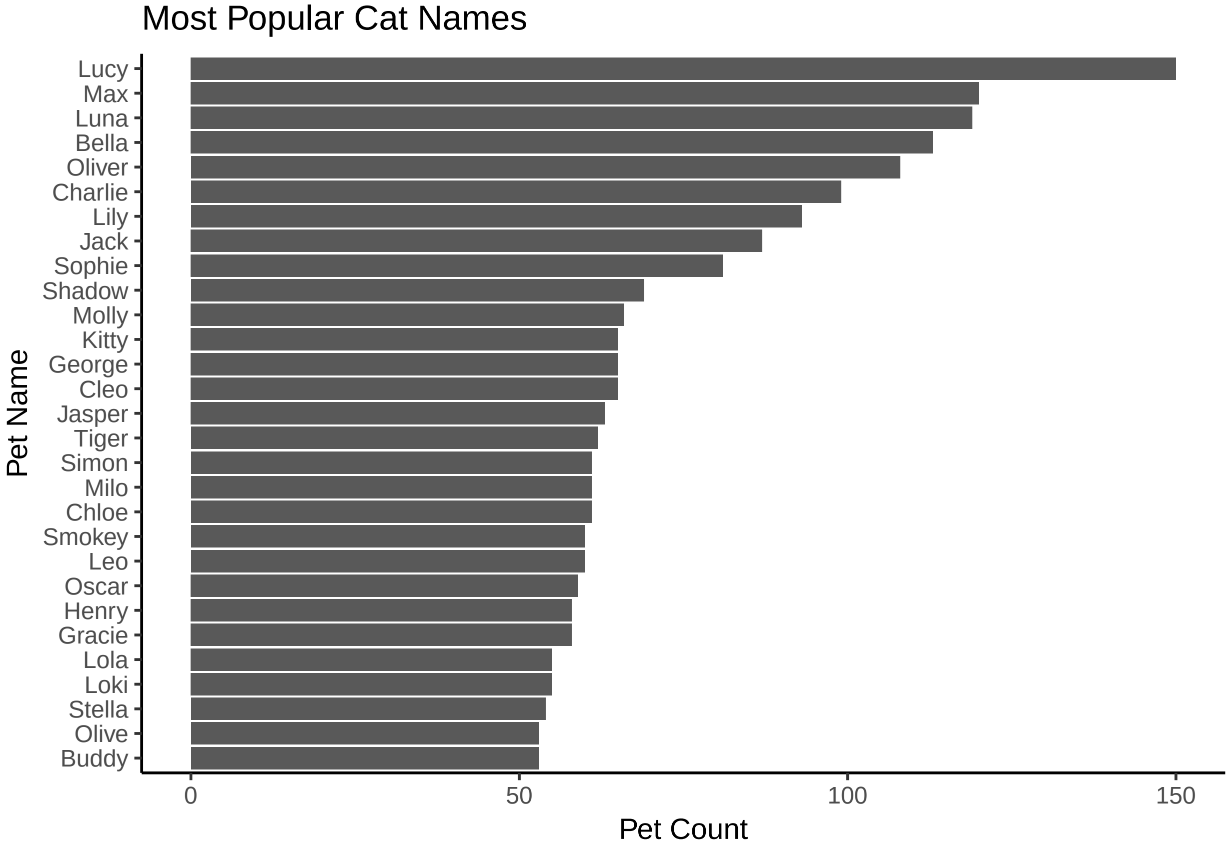
It appears that the highest concentration of pets is located in the North Seattle area.

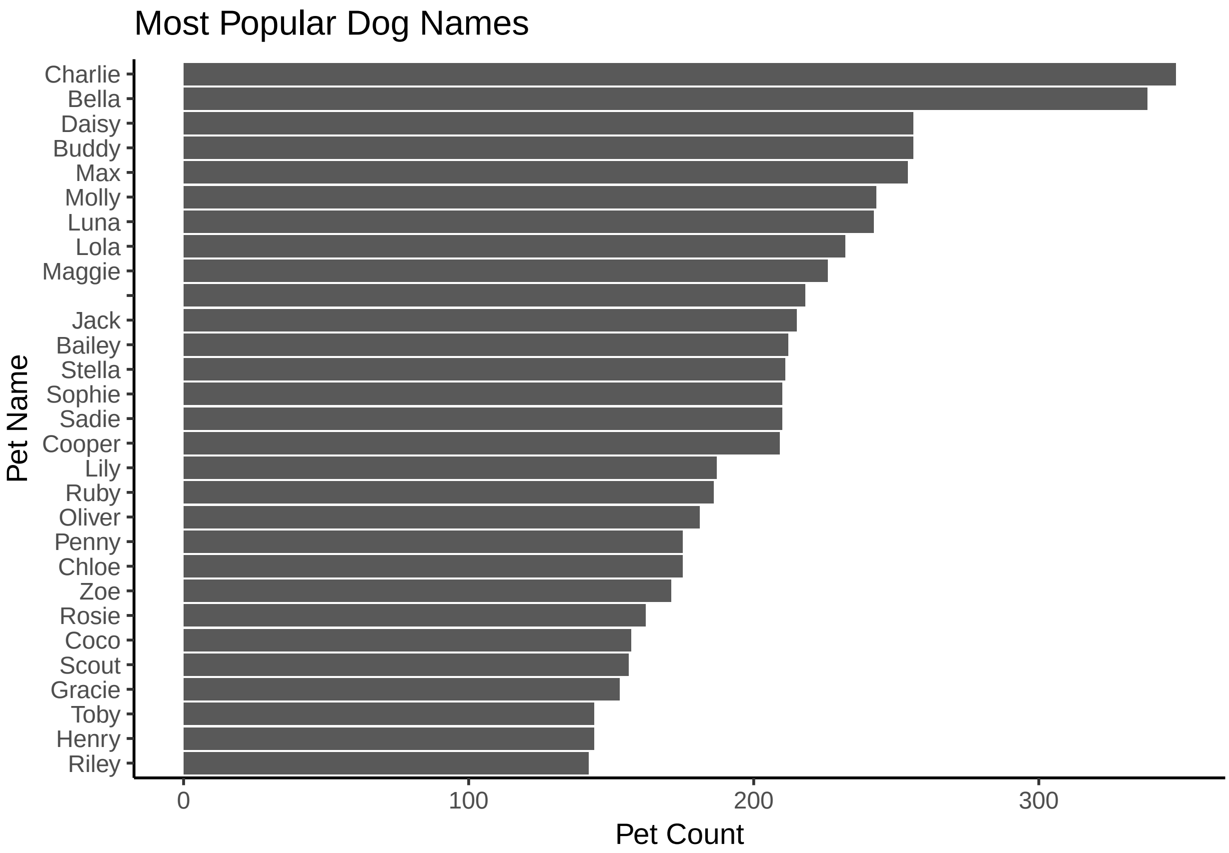
What's in a name?

This data set contained pet names which I found to be an entertaining data point. It enables us to answer the questions such as "What's the most popular dog/cat name in Seattle?" or "Is there any overlap between dog and cat names?"




Most Popular Names by Species
Cat
Dog
Livestock
Pet Name
Lucy
Bella
Charlie
Max
Luna
Molly
Buddy
Jack
Daisy
Sophie
Oliver
Lola
Lily
Maggie
Stella
Sadie
Bailey
Ruby
Chloe
Cooper
Zoe
Penny
Gracie
Henry
Rosie
Oscar
Coco
Scout
Milo
0
100
200
300
400
0
100
200
300
400
0
100
200
300
400
Pet Count


Most Popular Cat Names
Pet Name
Lucy
Max
Luna
Bella
Oliver
Charlie
Lily
Jack
Sophie
Shadow
Molly
Kitty
George
Cleo
Jasper
Tiger
Simon
Milo
Chloe
Smokey
Leo
Oscar
Henry
Gracie
Lola
Loki
Stella
Olive
Buddy
0
50
100
150
Pet Count



Let's see what we've learned!

- The most popular names overall are Lucy, Bella, and Charlie.
- Lucy is a common name for both dogs and cats.
- Lucy, Max, and Luna are the most popular cat names.
- Charlie, Bella, and Daisy are the most popular dog names.

Conclusion

We've done some high level exploration of our data set including:

- What's the range of data available?
- Where are pets concentrated in Seattle and does this vary?
- What are the most common pet names?

That's all for now!