# DS 100 Discussion 2 - Intro to Pandas

DS 100 Staff

January 2017

## 1 Intro

We've tried to make this discussion section a little bit different than a traditional worksheet - in addition to this worksheet, we've provided a bit of data to work with, under the disc02 folder in the github repository.

## 2 Handling missing values

Before we can work with our data, we need to deal with missing values. There are a number of ways we can do this.

- If we simply drop all missing values, how many dogs will remain? How can we perform this operation with pandas?
- How could we fill missing numerical values in order to keep some record of each dog?
- How could we fill missing "fur" values? There are multiple good answers.

## 3 Counting categorical data

- What's the most common type of fur?

- Do any two dogs share the same name?

## 4 Filtering and grouping data

- What's the average height of dogs with short fur?
- Some groups may have have disagreed on the exact value for the previous question. How could two different values be "correct"? What differing methodologies would lead to these two values?
- Dog height varies by furriness. What type of dogs tend to be tallest? You can find this in one line (i.e., not by filtering for each type of fur).

## 5 More advanced - Powerful Puppers

- Let's say that dog cuteness can be described by a simple formula:

  $$Cuteness = \frac{Age}{Height}$$

  How do the different varieties of dogs stack up to one another, in terms of average cuteness? What about in terms of maximum cuteness?
- Which is the cutest dog in each category of fur?

## 6 Vectorizing code - Big Doggy Data

Finally, let's look at how we can vectorize code in pandas, by converting some basic python code into its vectorized pandas equivalent. You should be able to perform the same operation as the given cell using only a single line of code operating on the clifford_data dataframe. We used the pandas

```
where, apply, and groupby
```

operators in our solution.