The Giant Database of Dogs:

Quantitative Literacy with a Furry Face

Leanne C. Powner, PhD
Leanne C. Powner Coaching and Editing
leanne@leannecpowner.com

This Draft: 28 Aug 2025
Dataset version: 2.0

The Giant Database of Dogs: Quantitative Literacy with a Furry Face

Abstract

This short paper introduces a data set of dogs constructed explicitly for pedagogical purposes. The simple cross-sectional structure and easily understandable variables allow students from any academic background to analyze the data. No substantive background is necessary; intuitive guesses allow students to hypothesize without any prior knowledge. Moreover, the regular updating process allows students to contribute to the data, as well as faculty to use the same assignment each term with results that will differ to deter plagiarism. The data are suitable for analysis in MS Excel as well as in any standard statistical program.

Introduction

One of the most tedious parts of preparing a methods course with a quantitative component is designing analysis tasks for assessment. The limited options created by the data sets which accompany most quantitative methods textbooks become boring, and most also require some substantive background knowledge on the student's part to propose hypotheses. In this paper, I present a simple crowdsourced project, The Giant Database of Dogs, for instructors to use as a supplement to existing classroom data sources. I describe the data set's properties, identify some sample analysis questions that can be posed to the data at varying levels of statistical complexity, and provide a student reproducible that describes the data collection methods and codebook.

Properties of The Giant Database of Dogs Data Set

This project builds on a similar idea used by Shane Gleason (Texas A&M University – Corpus Christi) in his own methods classes, where he has developed an in-house database of cats. I am a dog person, so I built a collection of dogs. Dogs also provide far more variation on many variables than cats, especially the critical continuous variable of weight. Version 2.0 of the data set, released with this paper, includes approximately 1,100 observations of unique dogs. This relatively large but still manageable number of observations allows for statistical analysis, even in MS Excel, that will produce statistically meaningful results.

The combination of a significantly larger dataset, the freely available nature of the dataset, and the rise of AI tools mean that most instructors will benefit from taking a random sample of observations and using only those in their own classrooms. This allows you to generate a 'new' dataset each term while still using the same assignment and script to generate your answer key.

Data collection was accomplished via a combination of convenience sampling and snowball sampling. I used my professional BlueSky, Threads, and LinkedIn, and personal Facebook, accounts to post calls for submissions, which were collected using a simple Google form. Some respondents on both platforms reposted the call to their personal feeds, and in some cases posted in Facebook groups of which they were a member. In other words, this is not a deliberately representative sample of dogs. You may wish to discuss with your students the implications of this for analysis, depending on the level of sophistication in quantitative analysis you are targeting.

The data set includes variables at all common levels of measurement. Nominal variables include the dog's name, primary color, gender, and status of being spayed or neutered. Ordinal variables include a five-value "size" indicator, ranging from toy to humungous. Values for all these variables are presented in text form, giving students an opportunity to learn to recode strings into numeric values if needed. Analysis in MS Excel will work with text values, though not cleanly as the

categories are not alphabetical;[1] SPSS and other statistical software will allow labeling to create an effective numeric code.

I include two continuous indicators, weight and age in years. The precision of age in years varies by respondent; you may choose to have students round before analyzing. Weight is expressed in pounds; those in metric-system-employing countries may choose to convert to kilos by dividing by 2.2.[2] Dogs in this sample (v. 1.0) range from 1.4 to 178 pounds (0.64 to 80.9 kilos). Breed status – one, designer/deliberate cross, and multiple – can be treated as a three-value nominal or ordinal variable, or as a count. The conceptualization of what this indicator represents, and therefore what level of measurement it is, would be a good discussion topic in your treatment of measurement.

New to the 2.0 release is a question asking whether your pup is a good pup. Obviously, the only acceptable answer to this question is yes. The data reflect this and therefore exhibit no variation (i.e., this is a constant). This question was introduced when approximately 1,000 observations had already been recorded; as a result, only about 100 observations have data for this variable.

Suggestions for Classroom Use

New versions will be released before each fall semester, including any new dogs that have been added since the previous release. This way, assignments can remain the same, but the answers will change so that students cannot reuse responses from previous terms. To encourage more data collection and to let students experience the form, I suggest asking each student to add one observation – either a dog of their own or one owned by a friend or family member who is willing to provide information. Having to decide how to handle some of the questions for their own case will help them intuit some of the (deliberate) shortcomings of this data collection form.

Many of the variables in the data set, including pet name, are string variables. Some software will calculate the modal value of string variables, allowing you to ask things like "what is the most common dog name" or "what is the most common name among female dogs." The latter requires sorting or conditioning on *gender*. MS Excel will perform this analysis using simple sorting tools and the COUNT command for strings. (If you prefer, you can generate de-stringed numeric codes in the copy of the dataset that you provide to your students for non-text variables so they can skip this step if this skill is not part of your course objectives.) Handling spelling variants of names could be a good discussion – should Daisy, Daisie, and Day-Z be pooled for analysis of most common name? Variants of Lily, Lucy, Lizzie, and other names exist as well, all apparently female.

While no cleaning should be necessary, you may want to have them investigate any potential duplicates. Checks for missing data show missing names and weights, but no other variables (no field was required by the survey). Is missingness of weight correlated with (or clustered in) a particular size? Decimal values for weight usually indicate conversion from weights provided in kilograms (approximately 10% of responses). *Weight* also shows significant clumping at round values, which should be evident in a histogram. Veterinarians consider *Age* less than 1 year old a puppy; ages over 7 are considered "senior" dogs.

The inclusion of categorial variables (and dichotomous variables especially) allows for the calculation of and tests involving proportions. Besides questions involving this data set, such as what proportion of dogs have some shade of brown as their primary color (thus requiring the pooling of the light and dark brown categories), future assignments could test whether the proportion of brown

[1] This is a good opportunity to introduce the IF command as a recoding technique.

[2] Respondents were given the opportunity to indicate the weight of their dogs in kilograms; I have already converted these to pounds. They account for most of the odd decimal values in the dataset.

dogs (or fixed dogs, or…) in the current classroom sample of the data set differs from a separate random sample.

Weight and age lend themselves to t-tests and potentially regression. The simplest question would be to ask if male dogs are on average heavier than female dogs. The next obvious step beyond that is to replicate the finding with a regression ($weight = b_0 + b_1 sex$), where we expect a negative sign on a sex variable coded in the conventional manner as 1 = female. Students might rightly object, however, that fixing a dog (spaying or neutering) would affect weight as well, suggesting a second variable to include. This is a simple introduction to the notion of control, where we get different coefficients on sex if we analyze *fixed* = 1 and *fixed* = 0 separately, and different coefficients on fixed if we analyze *gender* = 1 and *gender* = 0 separately. The combined reported results thus represent weighted averages of the two separate coefficients. A more complex analysis of control using *weight* and *size* is presented in my Appendix B of my *Empirical Research and Writing: A Political Science Student's Practical Guide* (2e, Cambridge, early 2026).

The intention behind separating dog *gender* and *fixed* status was to allow for students to create a new four-value variable based on the joint values. These could also be used to create an interaction term for regression analysis of weight, or even age. (Conventional veterinary belief is that fixing a pet prolongs its life in most cases.) This is an exercise in recoding, either in an if-then manner or through multiplication. The use of two dichotomous variables makes for easy interpretation of the resultant interaction term, though the direction of the hypothesis is not very clear.

The final question in the data set asks about the dog's *size*. While toy, small, medium, and large are standard terms for describing the size of dogs, the last category is usually "extra large." That said, while the terminology is relatively standard, the category breaks between them are not, and respondents were not given any guidance on classification. If we use the dosage for a common heartworm medication (Heartgard), small is 1-25 lbs, medium is 26-50, and large is 51-100 lbs; the drug does not come in standard doses above that. "Toy" is not a category, but few veterinarians would give a 5-10 lb dog (the lightest non-puppy dog [age > 1 year] in the data set is 4 lbs) the same dose as a 25 lb dog. This raises all sorts of interesting questions about self-classification.

It also provides an opportunity to explore identifying cut-points in the data to convert the continuous variable of *weight* into an ordinal one. In theory, the groups should be associated with relatively clear cuts or bins; are there dogs who appear to be misclassified? Say, Buster (observation 243), a 178-lb dog classified as large when all the other dogs around that weight and above are classified as humungous? Observation 49, Piper, is also identified by the respondent as being "large" but has a reported weight of 1.4 lbs; 30-lb Rikka (observation 13) is classified as humungous at age 14. Should these observations be dropped, since they're clearly incorrect in some manner? Or is Piper possibly a case of a puppy of a large breed? Respondents were given no instructions about how to handle puppies – code for expected size based on breed, or code based on current size? A Bernese Mountain dog puppy may currently only be 6-7 lbs and the size of an adult sneaker at two months, but it won't stay that way for long. Observation 425 may be a good example of this size-weight mismatch.[3] Discussions of how to improve the form with your students could lead to them deploying their own version of the data collection effort on your campus and compare if the consistency of responses is improved by the provision of more detailed directions.

Because statistics classes, or the days devoted to quantitative analysis in a more general methods class, can be stressful to students, the Instructor Packet also includes a collection of donated dog photos tagged with name, age, and weight. These are shared with permission and make an excellent addition to lecture slides to help decrease student tension.

[3] For those unfamiliar with dogs, Roxy is too young to be fixed, which happens between 4-6 months (0.33-0.5 years) for females.

Conclusion

The Giant Database of Dogs is designed to provide students with experience analyzing a wide variety of data types using simple tools that are commonly taught in introductory statistics or research design classes. The data set's theme of dogs – a common-knowledge topic requiring no special background – allows all students, regardless of departmental affiliation, to make reasonable hypotheses about relationships between variables. Consisting of a mix of nominal, ordinal, and interval-ratio data types, and with some obvious problematic design issues, it also allows students to both conduct analysis and discuss how the design issues affect their analysis. This is a key difference from other data sources, which are generally high-quality surveys and published data sets, where students do not get to critique the data collection instrument or process. With a new data set version scheduled to be released every year, and a dataset large enough to sample for classroom use, faculty will be able to reuse the same assignment questions, but students will obtain different answers, thus reducing faculty workload.

The Giant Database of Dogs

User's Guide, version 2.0
Accompanies dataset release 2.0

The Giant Database of Dogs is a crowdsourced initiative organized by Prof. Leanne Powner. The dataset is explicitly designed for classroom use.

Data Collection Procedures
Respondents were solicited by posts on Powner's professional social media profiles[4] as well as her personal Facebook profile and in two Facebook groups for political science data and faculty. Multiple posts occurred in both locations; Facebook included extensive review of Friends to tag as many known current or former dog owners as possible. Some respondents on both platforms voluntarily reposted the solicitation on their own feeds and/or in relevant groups to which they had access.

Respondents were directed to a Google form featuring a photo of Powner's current pair of beagles (see Figure 1). Instructions informed respondents that the data was for classroom and educational purposes only, and that no identifying information was being collected. As this research did not involve human subjects or animal testing, and additionally would be exempt as it was for classroom purposes, no human subjects clearance was sought. The Google form was configured not to collect any personally identifying information beyond the minimum needed for successful sending and completion of the form (i.e., an IP address, which it does not save after the form is complete).

Respondents were informed that they could submit multiple dogs and that both current and former dogs were welcome. "Best guess" data was acceptable where precise information was not known; this most likely applies to weight given the clustering of data at 'round' numbers (ending in 5 or 0).

The data set will be updated each semester with new inputs. You are welcome to add your own dog(s) to the data set using the form located at https://forms.gle/aDpw8b8tnUoYGgdU9 .

Data Cleaning
Data collection was closed when the sample reached 1,103 dog entries. Powner then minimally cleaned the data by assigning a unique identifier value (*Obs*) to each case, converting all weights to pounds (multiplying kilogram responses by 2.2), removing all text in the *Weight* and *Age* columns to leave only digits, and removing additional text such as nicknames provided in the *Name* field.

Codebook
Obs – unique observation number for case identification
Name – dog's name as provided by the owner (excess text removed)
Gender – {male, female} closed choice
Fixed – "yes" if spayed or neutered, as appropriate; "no" if not
Color – respondents selected one from {light brown, dark brown, black, white, reddish, yellow/blond, no primary color, other}. Other was not an open-ended response.
Heritage – 1 = single breed, 2 = designer/deliberate mix, 3 = mixed/unknown
Age – age in years, as provided by owner; converted if provided in months or fractions of a year.
Weight – approximate weight in pounds, as reported by owner. To convert to kilos, divide by 2.2.
Size – Toy, Small, Medium, Large, Humungous
GoodPup – is your pup a good pup? {yes, no} (only respondents Obs > 1000)

[4] A similar solicitation on LinkedIn generated only 239 impressions and is unlikely to have driven much traffic.

Figure 1 Data Collection Instrument. Response options collapsed for space reasons.

The Giant Database of Dogs

Building off an idea from Shane Gleason (TAMU-Corpus Christi), I'm building a giant database of dogs to use for classroom statistics exercises. Current or past dogs are welcome. Please provide your best guess for each of the questions below; do a separate form for each dog. The dataset will be made available for public instructional use after it reaches a sufficiently large N for analysis and I have a chance to clean it. Please feel free to share widely!

The data collected will only be used for classroom teaching and assessment purposes. Faculty in a variety of disciplines and institutions will create questions or assignments using this data, which contains no personally identifying data. Students may answer questions like "What is the most common male dog name? Are male dogs heavier on average than female dogs? Does dog size or being fixed (spayed/neutered) affect this conclusion?" All of these can be answered using concepts taught in introductory statistics and research design classes.

What is your good pup's name? [text response]

What gender is your dog? {male, female}

Has your pup been fixed (spayed or neutered, as appropriate)? {yes, no}

What is the primary color of your dog? {light brown, dark brown, black, white, reddish, yellow/blond, no primary color, other [closed response]}

What is your pup's heritage? {Single breed, designer/deliberate mix (e.g., labradoodles), mixed breed/unknown}

How old is/was your pet? [text response]

What is your pup's approximate weight? Please indicate pounds or kilos. [text response]

How big is your dog? {toy, small, medium, large, humungous}

Is your pup a good boy or good girl? {yes, no (seriously, why are you choosing this?)}

