Practice Exercises for Week 2: Data Visualization
in Tableau

This study guide will direct you through some of the analyses in our analysis plan that you can run on the "dognition_data_aggregated_by_dogid" data set using the Tableau techniques you learned this week. Please refer to Dognition Data Set Description document for a description of the data set, and download the "dognition_data_aggregated_by_dogid" data set from Week 2 to use with these exercises. These can be found in the Dognition Data Set, Description, and Analysis Plan reading in Week 2.

In the text below, all variable names will be depicted in *italics* and all properties on the Marks Card will be depicted in **bold**.

If you encounter any problems while working through these exercises, post your questions to the Module Forum Data Visualization with Tableau and don't be afraid to ask your peers for help!

Note that answers to these exercises will be needed to complete the graded quiz for Week 2, so it is important to complete all of the exercises!

Exercise 1

Before starting to address analysis questions, it's a good idea to get a feel for what our data look like. To get started, I suggest you use the Row and Column shelves and the Marks Card in the Tableau workspace to determine:

1) How many unique dogs are in the data set (indicated by the number of unique entries in *Dog ID*)?   17986
2) How many unique (human) users are in the data set (indicated by the number of unique entries in *User ID*)?   16261

As a discussion point, what does it mean if the number of unique *User IDs* is different from the number of unique *Dog IDs*?   some users might have more than 1 dog

Another thing it is helpful to do before addressing our primary analysis questions is to examine whether there are any questionable values in your key fields that may be mistakes or outliers. I suggest that you closely examine all of your variables, but I will highlight a few in particular here.

First, place *Country* in the Rows shelf as a dimension to see what countries are represented in our data set. If you turn that *Country* pill into a measure, you can aggregate it by Count (Distinct) to see how many countries there are in total.

Next, do the same for *State*. When you do, you will find that many entries in *State* have curious values! This is an example of how real-world data sets can be messy. To get an idea of what some of the strange values might mean, place *State* on the Rows shelf and *State* aggregated by Count on **Text** of the Marks card. You should get a table with all the possible values of State in the first column and the number of rows of data that have these values in the second column. Right-click (or CTRL-click) on a row in the table and select "View Data" to see what the raw data underlying each row in your table look like. Any ideas of what the numerical entries in the *State* field were likely supposed to represent (at one point)?

As another quality check, place *Sign in Count* on the Rows shelf. Then de-aggregate the data in the workspace through the "Analysis" menu. You will see that there are some clear outliers in this variable. If you right-click/CTRL-click on the individual data points to see the raw data underlying them, note what columns of data seem to be common to all the rows displayed at this point. We will want to examine these outliers and possibly exclude them from future analyses.

Recall that one way to exclude outliers is to group together the individual *Dog IDs* that are associated with all the extreme points on the graph, so that they can be filtered out of your analyses. For that to be possible, you will need to put the *Dog ID* field on **Details** of the Marks card. Note the difference between the raw data underlying the points you click on now, compared to before you put *Dog ID* on **Details**. Group together all the data points that have *Sign In Counts* of above 175.

We have confirmed that these Dog IDs represent test accounts that Dognition used to troubleshoot their website. For the rest of the practice exercises using the dognition_data_aggregated_by_dogid data set, exclude these *Dog IDs* from your analyses by putting the grouped variable you just made on the *Filter* shelf and excluding all of the data points in the group with extreme values.

Exercise 2

Now that we have a feeling for what our data look like, we can start addressing some of the questions in our analysis plan. Our job is to make a recommendation to Dognition about what they could do to increase the number of tests customers complete. One way to start making hypotheses about the business changes that would be useful, is to identify the features of dogs or their owners that have correlated with increased completion scores in the past.

The dognition_data_aggregated_by_dogid data set is well-suited for those types of questions, because it provides data about the total number of tests completed by each dog and/or human customer (as opposed to data about each test completed, which is a data set we will analyze next week).

In this exercise, we will focus on examining features of dogs that correlate with differences in the number of tests the dogs complete. The field representing the total number of tests customers completed is *Total Tests Completed*. We have also computed the mean and median "ITI," or "inter-test interval," in minutes and hours for the tests each dog completed, as well as the amount of time that elapsed between the first and last tests each dog completed.

To address the question of how aspects of dogs' personalities and breed types affect completion metrics, you can make visualizations with any of the variables described above, and the *Breed Type*, *Breed Group*, or *Dimension* variables. Try different combinations of dependent variables and independent variables in the Rows and Columns shelves and/or the Marks card. To aid in your interpretations, it may also be useful to know whether certain types of dog breeds tend to have certain types of personality types.

sporting dog->charmer and socialite.

As you complete these visualizations, recall that the *Dimension* field represents personality dimensions that can only be assigned to dogs after they have completed the Dognition Assessment, which is the first 20 tests.

Overall, what trends do you see? Make notes for yourself on this.

Exercise 3

In this exercise, we will focus on examining variables that might give insight into owners' personalities. We will start with the *DNA tested* and *Dog Fixed* variables.

Owners who get their dogs DNA-tested are likely to be doing so (a) to confirm whether or not the dog is pure-bred, or (b) because they are really interested in finding out more about where their dog came from (http://www.caninejournal.com/dna-testing-for-dogs/). If either of these motivations are true, you might infer that such owners are inherently interested in learning more about their dog, and therefore hypothesize that these owners will be more likely to complete Dognition tests than other customer groups, due to the new insight the tests will provide about their dogs.

In addition, you might hypothesize that these effects would interact with what kind of breed type (pure-bred, mixed or unknown origin, etc.) the dog is, and whether or not the owner intended to breed the dog (which would be impossible if the dog were fixed, or spayed/neutered). Owners who intend to breed their dogs might be more interested in the personality of their dogs than others (because they will be interested in whether they are breeding a certain kind of temperament).

If you disagree with these hypotheses, write down your own hypotheses! Then test them by making visualizations with *Total Tests Completed*, *Breed Type*, *Dog Fixed*, *DNA Tested,* and *Total Tests Completed*. Remember to make sure each variable is correctly classified as a measure or dimension, as appropriate.

Exercise 4

In this exercise we are going to address the question of whether particular countries and/or states within the US have more Dognition customers than others. We will also address whether customers within specific geographic regions are more likely to complete more tests.

First, let's find out what countries Dognition users tend to come from. To do this, we will ask how many distinct dogs have been tested in each country. Start by putting *Dog ID* aggregated by Count (Distinct) on the Rows shelf, and *Country* on the Columns shelf. Tableau will likely give you a bar graph (unless you started by using another type of graph).

There are a lot of countries and it's hard to get a good idea of the results when there is so much to scroll through, so let's use a Tableau's map instead. Click on the filled map icon in the *Show Me* card. Now each country should be colored by how many dogs have been tested there. This graph will illustrate clearly that most customers are located in the United States. To dig into this further, use the Filter shelf to limit what you are visualizing to only US records. Are there any states that clearly have more customers than others?

Next let's examine whether there are any states that have customers that are more or less likely to complete their tests. Repeat the graphs above using *Total Tests Completed* as your dependent variable. What patterns do you see? As you interpret those patterns, make sure to also check how many records are being aggregated in each state.

One thing that becomes immediately apparent as you make these graphs is that the median and average numbers of tests completed seem to be higher than what we've seen in a lot of previous graphs. Is it possible that US customers tend to finish more tests than non-US customers?

Use *Country* to make 2 groups: the first group should consist of the US, and the second group should consist of all other countries. Now try some visualizations with *Total Tests Completed* and your new *Country (grouped)* variable. How would you make a visualization that would let you test if there are any countries who tend to complete more tests than US customers AND who have enough records in them that you would feel comfortable interpreting their results?

As a point of discussion, how can you use the information from these geographic analyses to make a recommendation about a geographical market Dognition should either nurture or move more into?

Exercise 5

In this exercise, we are going to try to assess the hypothesis that customers who complete tests quickly are more likely to complete more tests overall. As often happens in real-life scenarios, the data we have in this version of the data set are not perfectly suited to answer this question. Nonetheless, we can still gain some useful insight.

The two variables we have available related to test completion rate are *Median ITI* and *Mean ITI*. It might be helpful to review how these fields were computed. I took all the time-stamps (i.e., the recorded day/time of completion) from the test records of a given *Dog ID* in a separate data set (which we will use next week), and computed the amount of time between the time-stamp of each test. That gave me a collection of inter-test intervals for each dog.

If a dog completed 4 tests, I recorded the mean and median of 3 inter-test intervals. If a dog completed 20 tests, I recorded the mean and median of 19 inter-test intervals. An important fact to remember here is that customers could complete the first 20 tests of the Dognition Assessment as quickly as they wanted, but then they would only be able to complete one test per month after that. Therefore, our question about completion rate really only makes sense if we look at the first 20 tests.

However, since the dognition_data_aggregated_by_dogid data set has the data aggregated at the level of *Dog ID*, not at the level of individual tests, we can't just filter out all the data associated with tests 20-45. Instead we have to filter out all the data from dogs who completed 20 or more tests. Sadly this means (a) we will throw out some data in this analysis that could have otherwise been useful, and (b) our analyses will be biased towards dogs who do not complete all the Dognition Assessment tests, which might not translate well to dogs who do finish the Dognition Assessment tests. Such is life, in the data world.

With all of that in mind, we can try to address our question about completion rate by running a set of regressions treating *Total Tests Completed* as a measure. Start by filtering out all of the data from dogs who completed 20 or more tests. Then put *Total Tests Completed* on the Columns shelf. Then put *Median ITI* and *Mean ITI* on the Rows shelf, and un-aggregate the data. After that's completed, add a trend line to each graph. You might be surprised to observe that the trend lines appear to go in opposite directions. How can that be? I suggest you watch the video we supplied in the course materials about the effects of outliers again. You may also find it helpful to look over these resources about how outliers can affect regression lines:

http://stattrek.com/regression/influential-points.aspx?Tutorial=AP
http://discuss.analyticsvidhya.com/t/effects-of-outliers-on-regression-model/2403/2

What do you think the best interpretation of these results are? Would these results influence any recommendations you make to Dognition? Use the Course Discussion Forums to ask any questions or share your thoughts.