**Tableau Team Assignment 1**

In the next two group exercises, we will use the Dognition data set to practice using visualization to answer analysis questions. Please watch the “Meet Your Dognition Data” video to review the business problem Dognition would like to solve. Download the “dognition\_data\_aggregated\_by\_dogid” data set for this assignment, and the “Dognition Data Set Description” document for a description of the fields that you may see. These exercises will direct you through some of the analyses outlined in the Dognition analysis plan (provided on Sakai; the format for this analysis plan will be explained later in the course) that you can run on the data set using Tableau.

Please submit one document per team with your team’s answers to each question in this assignment through Sakai. Make sure all of the team members are listed on the top of your submission.

In the text below, all variable names will be depicted in *italics* and all properties on the Marks Card will be depicted in **bold**. All the questions in this Assignment can be answered using the Row shelf, the Column shelf, and the Marks Card in Tableau.

If you encounter problems while working through these exercises that you and your team can not solve on your own, you can:

* Review the Tableau videos from this week
* Use resources that you find online about how to make specific types of graphs in Tableau
* Ask questions in the “team\_assignment\_1” folder in Piazza
* Ask the TAs for help

The answers to these exercises will serve as the basis for Quiz 1. Although you should submit your answers to these exercises as a team, **you must take the quiz on your own without any help from anybody else.**

**Exercise 1**

Before starting to address analysis questions, it’s always a good idea to get a feel for what your data look like.

1. Start by determining the number of unique dogs and number of unique human users in this data set. What data fields did you use to answer that question, and why are the numbers different?

Another thing it is helpful to do before addressing your 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.

1. How many unique values are there in *State*, not including null values?
2. Make a table in the Tableau workspace 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. Take a screen shot of how you did this (make sure all the shelves and cards are included in the screen shot) and paste the screen shot in your assignment submission.

You should have seen that many entries in *State* have curious values. This is an example of how real-world data sets can be messy! Look at the raw data underlying the rows that have numbers in the *State* field.

1. What do you think the numerical entries in the *State* field were likely supposed to represent when they were originally created or entered?

As another quality check, place *Sign in Count* on the Rows shelf. Then de-aggregate the data in the workspace. You will see that there are some clear outliers in this variable.

1. How many users signed in more than 175 times? Hint: Some users signed in the same number of times, so you will likely NOT get the right answer if you simply count the number of circles you see by eye in your workspace.

Look at all the data points that have values above 175. I have confirmed that all the Dog IDs associated with these rows of data (the ones that have signed in more than 175 times) represent test accounts that Dognition used to troubleshoot their website. 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. Make sure the group you end up with includes the correct data you want to exclude, and only the data you want to exclude. **For the rest of the exercises or quizzes using the dognition\_data\_aggregated\_by\_dogid data set, exclude these *Dog IDs* from your analyses using the *Filter* shelf. Unless stated otherwise, all exercises and quiz questions will assume these Dog IDs have been excluded**.

In addition, the Dognition data set has an “exclude” field built into it that indicates when rows of data should be excluded for other reasons determined by the Dognition team. **For the rest of the exercises or quizzes using this data set, exclude the *Dog IDs* that have a “1” in this field from your analyses using the *Filter* shelf. Unless stated otherwise, all exercises and quiz questions will assume these Dog IDs have been excluded**.

**Exercise 2**

Now that you have a feeling for what your data look like, you can start addressing some of the questions in the analysis plan. Your ultimate goal, remember, 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 in the next Group Assignment).

1. How many Unique Dog IDs are labeled as “Pure Breed”, “Popular Hybrid”, “Mixed Breed/Other/I Don’t Know”, or “Cross Breed”?

Start by 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*. I have also computed the mean and median “ITI,” or “inter-test interval,” in minutes and hours for the tests each dog completed for you, 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.

1. How do the completion rates of dogs with non-null and non-zero entries in the *Dimension* field compare to those who do not have entries in the *Dimension* field? Why?

Next, examine variables that might give insight into owners’ personalities. Start with the *DNA tested* and *Dog Fixed* variables.

Owners who get their dogs DNA-tested might 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).

1. How do the number of completed tests compare between DNA-tested and non-DNA tested dogs? What about the number of account sign-ins, and the amount of time between the first and last games? What do you think this means about DNA-testing as a way to provide insight about the human Dognition customers?

**Exercise 3**

Your next task is to examine whether particular countries and/or states within the US have more Dognition customers than others. You will also address whether customers within specific geographic regions are more likely to complete more tests. Begin by finding out what countries Dognition users tend to come from.

1. Make a map of the world that is color-coded by how many unique dogs have been tested there. Take a screen shot of how you did this (include all Tableau shelves and cards in your screen shot) and paste the screen shot in your assignment submission.
2. Make a map that is color-coded by how many unique human customers come from that country. Create a meaningful title for the map, output the map as a picture (title and caption included), and insert the resulting picture into your assignment submission.
3. What country has the most unique human customers, and how many “unknown” country assignments are there, according to the visualization you made in Q10?
4. There should be a country in Africa that has a deep color on your maps, suggesting it also has a lot of dog customers. What value was entered into the Country field for those data points? Why were they shown on the African continent?

Now examine whether there are any American states that have customers that are more or less likely to complete their tests. Filter the data so that you are only looking at data from the United States, repeat the graphs above using Total Tests Completed as your dependent variable, and make a note of what patterns you see. As you interpret those patterns, make sure to also check how many records are being aggregated in each state.

1. Make a map that shows the average Total Tests Completed for each American state, and that shows the number of records being aggregated into the data for each state. Restrict the field of view to the continent of North America only. Hover over a state, and take a screen shot that shows all of the items in your shelves or in your cards and that illustrates that the number of records indeed shows up in the tool tip when you hover over a state. Insert the screen shot into your assignment submission.
2. Make a visualization that illustrates whether there are countries whose customers tend to complete more tests than US customers AND who have enough records in them that you would feel comfortable interpreting their results. Output the visualization as a picture and insert it into your assignment submission. Explain why you chose the metrics or criteria that you did.
3. Based on your visualizations so far, what recommendations might you make about geographical markets Dognition should either nurture or consider focusing on, and why? Make sure you support your recommendations with data.

**Exercise 4**

Next, you should test 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 both the mean and median of its 3 inter-test intervals. If a dog completed 20 tests, I recorded both the mean and median of its 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, you can get preliminary insight into the question about how quickly customers complete tests and completion rate by running a set of regressions treating *Total Tests Completed* as a measure. Start by excluding all of the data from dogs who completed more than 20 tests. Then make 2 scatter plots in one Tableau workspace with *Total Tests Completed* on the x axis and *Median ITI (minutes)* and *Mean ITI* *(minutes)* on the y axis, and with all the *Dog IDs* as separate marks.

1. Fit a linear trend line to both plots. Take a screen shot that shows all of the items in your shelves or in your cards and insert the screen shot into your assignment submission.
2. What is the equation for the linear fit to the Mean ITI plot?
3. What is the equation for the linear fit to the Median ITI plot?
4. Why do you think the trend lines go in opposite directions?

Keeping all of the data from dogs who completed more than 20 tests filtered out, now plot the average of the mean ITIs and median of the median ITIs for all the dogs aggregated together, with *Total Tests Completed* on the x axis (so you should have 20 marks on each graph).

1. What value do you get for the average mean ITI for dogs who completed 8 tests? What value do you get for the median of median ITIs for dogs who completed 8 tests?
2. Would the trends you’ve seen in the ITI plots influence your recommendations to Dognition? If yes, describe how and why they would influence your recommendations. If no, describe why they wouldn’t influence your recommendations.