Project Journal

I. Project Goal

My project goal is make a recommendation to Dognition about what the company could do to increase the number of tests customers complete on its website. One useful approach is to identify features of dogs or their owners that have correlated with increased completion scores in the past.

In this project, I use total tests completed by each dog or each user as a key indicator and focus on study how the company could do better in user retention.

II. Dataset Description

The dataset is collected through Dognition's website. The dataset has 17,986 unique dogs and 16,261 unique users. Every user has approximately one dog on average but some users have more than one dog. It has 30 variables as shown in Part II.

III. Check questionable data for mistakes, outliers and missing data (after remove data with variable Exclude=1, the dataset has 17,828 rows)

Dimension (Category):

No.	Variable Name	Mistakes	Missing Data	Note
1	Dog ID			
2	User ID			
3	Gender			
4	Breed		15	
5	Breed Type			
6	Breed Group			0? dog not in the 7 categories? Or missing
7	Dimension		13,678	0 means no dimension for the dog; only dog completed at least 20 games will be assigned one of nine categories of dimensions

8	Membership ID			0 means the user does not subscribe
9	City			
10	State			A state name only with number means its capital city of the country and no state name available
11	Zip			Some Zipcode are 0 or 25; not clear what the data represent
12	Country			
13	Free_Start_User			
14	Last Active At			
15	Membership Type			0 means the dog owner does not subscribe service
16	Subscribed			
17	Excluded			

Measures (Quantitative):

No.	Variable Name	Mistakes	Outliers	Missing Data	Note
1	Total Tests Completed				
2	Mean ITI (Days)			1,256	Those are the users who completed at most 1 game; hence, no time interval could be recorded between the first and the last game.
3	Mean ITI (Minutes)			1,256	
4	Median ITI (Days)			1,256	
5	Median ITI (Minutes)			1,256	
6	Time diff btw first and last games (Days)				0 for only complete at most 1 game

7	Time diff btw first and last games (Minutes)				
8	Weight				0 weight? Maybe less than 1 lb. or data are not available
9	Dog Fixed				
10	DNA Tested				
11	Sign in Count		1 data point >175		>175 are test accounts
12	Max Dogs				

IV. Exploratory Data Analysis

Dependent Variable: Total Test Completed

Mean	Median	SD	Distribution
9.780	7.000	7.786	Not Normal

Approximately 40% of dogs accumulatively completed between 0 to 5 tests, 20.49% completed 20-25 tests and 20.21% 5-10 tests overtime. 76.78% of dogs finished less than 20 games, for which there does not exist enough information to a dimension about its trait and hence those data are less valuable to the company.

I suggest to separate the data set into two groups to compare the feature differences: dogs completed less than 20 games and dog completed at least 20 games since it begun the first game on Dognition website.

V. Hypothesis Test

1. How aspects of dogs' features affect the completion metrics?

    i. Are dogs with certain dimension tends to complete more tests than others?

    Overall, dogs with each dimension have approximately same mean, median and standard deviation in total test completed. However, among all dogs completed at least 20 games, the dimension Socialite rank highest with 20.77% and is more likely to complete at least 20 games, followed by Charmer 16.57% percent. Only 2.97% of dog completing at least 20 games is Einstein. However, note that the difference may also occur if the dogs' distribution over the 9 personal dimension categories is unbalanced in the real world, which though I do not have data to test out.

    Hence, weather the dogs completed at least 20 games or not is dependent on its dimension if the hypothesis that the distribution of 9 personal dimensions for dogs is balanced is true.

    ii. Whether certain types of dog tend to have certain type of personality type?

    Give a dimension a dog, I can know the composition of the breed group of the dog and compared it with the composition of dog on its website. For example, Sporting dogs are more likely to have a personal dimension of socialite, compared to other dog breed groups.

2. Does the breed type of a dog associate with higher or lower total number of test completed?