# DataXperts Summary:Dognition Data Set

By Alexia Kim, Christina Castillo, Cristina Ortiz, Isabella Barboza,Vien Nguyen

Our team, DataXperts, will analyze the Data Set Dognition using Jupyter Notebooks as a tool and Python as a language. After exploring the Data Set for Dognition, a fun fact I noticed is a significant amount of undefined or unlisted data that is listed as a “0” or N/A. Undefined data can be found in the breed, weight, and geographical columns. I am excited to explore the undefined data, as missing breed and geographical information can negatively impact our breed-specific behaviors and geographical data analysis. Overall, I believe our group will need to be careful in handling undefined data and discuss the best way to deal with it, whether through filtering or a different method, to ensure we get accurate analysis.(Cristina Ortiz)

Some of the tools provided that I found useful in analyzing the Dognition data set were SQL and Excel, which helped in organizing and arranging the data sets and provided the tools to identify any related sources that can increase the completed dognition tests related to the owners of tested dogs. After exploring the Dognition data sets, something I found interesting regarded the breeds of dogs being mixed. Some of the dogs with known mixed breeds are classified with either one breed or the other alphabetically, in order to not duplicate. Some dogs are just stated to be mixed and are part of a broader and undefinable group. I am interested in finding out how the behavior and test results performed by the mixed-dog breeds category compare to the behaviors and test scores gathered from other dog breeds and how their owners impact these scores. Our team will be using the Jupyter notebook and Python language to further analyze and communicate our team's findings regarding the questions provided for the Dognition data set. -Isabella Barboza

The analysis of Dognition user data highlights significant concerns regarding data integrity and the accurate tracking of user behavior, particularly between paying and promotional users. Key issues include the Free_Start_User column, which contains multiple data types: numeric values (0 for not free start users, 1 for free start users), nulls, and text of Canadian neighborhoods. This inconsistency complicates the understanding of user engagement and could misrepresent user statuses. Additionally, aggregated columns show NaN for null values, indicating potential gaps in capturing user status, while the Last_Active_At column exhibits mixed data types and includes invalid dates, hindering data validation efforts and oddly does not match up with the non-aggregated column.

The data was primarily analyzed using Python and SQL, with some outputs printed to Excel due to display limitations in Python. Despite the integrity issues, essential columns such as Created_at, User_id, and Test_name in the non-aggregated data were fully populated, facilitating effective analysis. However, an unusual finding was that the Created_at and Updated_at timestamps were identical for all entries except one, raising questions about the data collection process. **Overall, these findings suggest a need for improved data quality control measures to ensure accurate tracking of user behaviors and to mitigate misinterpretations of user engagement trends.** -*Christina Castillo*

At first, Dognition dataset requires some data cleaning before could use it to extract insights, especially because of many missing value and inconsistent in the “dog_id_max_ranks” sheet. I used tools like Python, SQL and Excel to sort, filter and validate the data for accurate analysis. One thing I found is that the Mean ITI and Median ITI which measure the time between tests are varied a lot between different dog breeds and user types. For example, dogs classified as “maverick” tend to have longer gap between tests, suggesting their owners might have more challenges staying engaged to the tests. Also, promotional membership type (type 4) looks like completed fewer tests with longer gaps between them, maybe due to less commitment.

Next I am planning to visualize the distribution of Inter-Test Interval (ITI) across different dog breeds using histogram, which will help identify trends in engagement for specific breeds or personality types. In addition, I will apply scatter plots to explore relationships between time gaps between tests and test completion rates across different personality types of breeds. I believe this will allow me to look for any correlation between engagement levels and personality dimensions. I will also use bar-chart to compare average ITI for different breeds or personality types, which will provide a clearer view of how engagement varies across groups. My goal for all of that is to explore whether we can conclude or predict user engagement levels based on personality dimensions and time gaps between tests. (Vien Nguyen)

While I was working with the dognition dataset (specifically the aggregated dataset), there are two columns for time difference between first and last game, one in days and one in minutes. A fun fact I discovered is that the mean and median columns for ITI (in days and in minutes) are calculated from the values of total time difference and the total tests completed. What I noticed while I was working with the dataset was that the values for time difference in days and in minutes that show up as “0” have unusual values for the Mean and median ITI. If there are “0” values for time difference, the mean and median ITI in days shows up “NaN’ and the mean and median ITI in minutes show up as “#VALUE!”. The approach to eliminating these undefined/missing values could be very challenging. However, Python is the ideal coding tool to utilize for deleting unnecessary data from excel files to display graphs of the correlation between total tests completed and the time difference between first and last game. (Alexia Kim)