Initial Data Exploration

Predicting Mastery in Academic Assessments

All work described in this report can be viewed in the initial_data_exploration notebook.

Download the **grockit_all_data.zip** file from the What Do You Know? competition on Kaggle. Import and inspect raw training data. Isolate relevant variables, fill or calculate new variables, and organize the dataframe. Resolve missing, invalid, corrupted or duplicate values.

Import Packages, Dataset & Inspect Raw Data

The raw training dataset contains 4,851,475 rows, or observations. Each observation holds information about a question attempted by a user. There are 17 columns, or variables. Of particular interest are **user_id** which identifies individual users and **question_id** which identifies unique questions. There are 179,106 users and 6,045 unique questions in the raw dataset.

Organize Columns

Redundant columns such as **outcome, answer_id, question_set_id** were removed. Variables related to time were converted to datetime objects.

A timedate column was created: `round_ended_at`. The **round_ended_at** column was filled with `answered_at` values if present. If no value was found in any `answered_at` field, then the value from `deactivated_at` was used to populate `round_ended_at`.

Evaluate & Resolve Missing Data

NULL values in the dataframe indicate that the value is missing, invalid or corrupted. Looking at the count of NULL values for every column in the dataframe revealed that timedate variables contained missing data. NULL values were expected for **answered_at** since this denotes that the question timed-out. Missing values for **date_of_test** were also no surprise since it was known that users may or may not enter an expected test date. NULL values for **round_started_at** and **deactivated_at** were less understood and, therefore, examined more closely.

All 69 observations with missing time data for all timedate variables were eliminated. Another 21 rows were dropped for missing **answered_at** and **deactivated_at** values since round duration cannot be calculated in those rows.

3,824,208 of the 4,851,475 observations in the raw dataset were missing values for **date_of_test**. That is nearly 80%, but this does not indicate how many users entered no test date data. A look at NULL values by user revealed that 87% of users did not provide an expected test date. This variable cannot be adequately analyzed so the column was dropped.

At this point, most of the obvious NULL values in the dataframe were in the **`answered_at`** column and 10 remained in the **`deactivated_at`** column. Since observations where both are NULL were eliminated, there was no harm in keeping all of the remaining NULL values. Either variable provides the necessary information for `round_ended_at`, so ignoring one missing value is no problem.

Not at all obvious from traditional inspection of the dataframe was the fact that NULL values existed in the **`game_type`** column. This was learned from looking at the labels in the `category_labels.csv` file. Information found in the forum where the datasets are hosted claim a value of 6 represents a data collection failure. Upon inspecting observations for these values, they were dropped mainly because most, if not all, of the observations involve rounds with end times earlier than start times.

Evaluate & Resolve Duplicate Data

Since one user cannot physically start or end a round at the exact same time more than once, records were inspected for such a scenario. This revealed many duplicates, but it also revealed many observations where `round_started_at` times were the same as `round_ended_at` times. Given the reasonable expectation that it is impossible to start and end a round at the same time, 4264 observations meeting such conditions were dropped.

After weird time issues were resolved, duplicate observations for the combination of **`user_id`**, **`round_started_at`** and **`round_ended_at`** were dropped.

Compute and Explore Time Data

The **`round_duration`** column was created and filled by subtracting `round_started_at` values from `round_ended_at` values.

A view of time data statistics revealed that some rounds were days long and many were over within only a few seconds. Rather than delete the 167,564 observations outside a threshold of 5 seconds to 10 minutes, new columns were created to categorize fast and slow extremes. One variable indicating the number of **`prior_attempts`** of a particular question by the user, and one variable indicating when these become **`excessive_attempts`** (> 100) were also created.

Before and After Data Wrangling:

	Total Observations	Total Variables	Number of Users	Number of Uniques Questions
Pre Cleaning	4,851,475	17	179,106	6,045
Post Cleaning	4,846,730	22	179,033	6,045
ELIMINATED	4,745	(9 added) 4	73	0

Variables related to user statistics were added near the end of the notebook. They are mainly for EDA purposes in the next notebook.