KKBox had provided all the necessary data. There are 4 datasets:

- Train dataset: information on whether or not a user has churned
- Members dataset: user information for all KKBox users
- Transactions dataset: monetary transactions for all KKBox users
- User Logs dataset: listening behaviors for all KKBox users

The train dataset contains churn information for 970,960 users. Fortunately, there are no missing values in this dataset, so there is nothing to clean. It is interesting to note that a majority of the users have renewed their subscription (is_churn = 0); only 8.9% have churned, meaning the dataset is pretty biased data set.

The rest of the datasets, however, needed to be cleaned. The major theme of this data wrangling was reformatting dates into datetime objects and converting certain columns into categorical variable. The dates need to be in datetime objects so it can be filtered according to its date elements. Leaving the columns as integer type instead of the category type can lead to confusion since the values don't actually represent the numerical value but are placeholders. Categorical variables take on a limited, and usually fixed, number of possible values.

The members dataset contains user information for 6,769,473, which is a lot more than needed. The only relevant user information are those that correspond to the user part of the dataset. Once dataset is filtered to only the relevant ones, the data is cut down to 860,967 which means that the dataset does not have user information for all the users in the train dataset. 'Registration_ini_time' contained data that should be in the datetime format. 'City', 'gender', and 'registered_via' were all converted to categorical variables. After creating a boxplot for 'bd' (aka age), it is apparent that this column contains many outliers. The minimum value was -7000 while the maximum value is 2016; these values do not make sense in terms of age. Outliers under 1 year old and over 100 years old were removed, so the data now only contains age information for 386,715 users. This removed more than half, but keeping those outliers would have really skewed the data.

Moving onto the transactions dataset, which contained 1,431,009 rows. A user can make multiple transactions, meaning there can be multiple rows per user in this dataset. After filtering the dataset to only those in the train data, the dataset now contains 1,132,036 rows. Values in the 'transaction_date' and 'membeship_expire_date' were converted to datetime objects. 'Payment_method_id', 'is_auto_renew', and 'is_cancel' were treated as categorical variables. It seemed odd that 'payment_plan_days', 'plan_list_price', and 'actual_amount_paid' had minimum as 0. However, after further investigation, it seems that it is possible for those columns to have a 0 value. This just means that either the user did not choose a payment plan, or the user received a free trial.

Lastly, the user logs dataset contained 18,396,362 rows. Each row represents user listening behavior at the day level; therefore, there are multiple rows per user. The more rows a user has means the more user uses the app. The dataset was filtered down to only contain those in the train set, so the dataset contains 13,532,944 rows. The only column fixed was the 'date' column, which was converted to a datetime object.