Cedric Herman
Mentor: Srdjan Santic
Career coach: Allison Matthews

Springboard
January 2018 cohort

Capstone #1: Data Wrangling

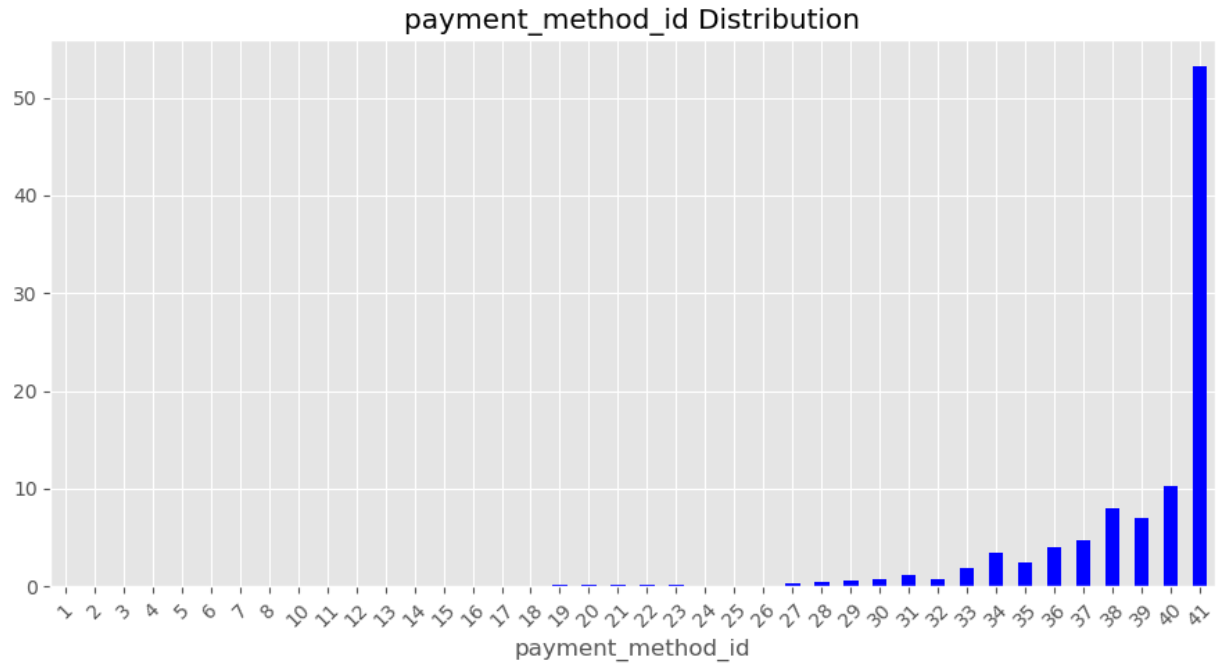
KKBOX (streaming music service provider) made available 3 data files to get to know their customer base behavior. Transaction CSV file contains transaction history from January 1st, 2015 to February 28th, 2017. Information about plans subscription and prices are included. User log CSV file captures user's daily activity such number of songs played from January 2015 to March 2017. And finally, member CSV file features demographic information on users who created an account on KKBOX.

Contents

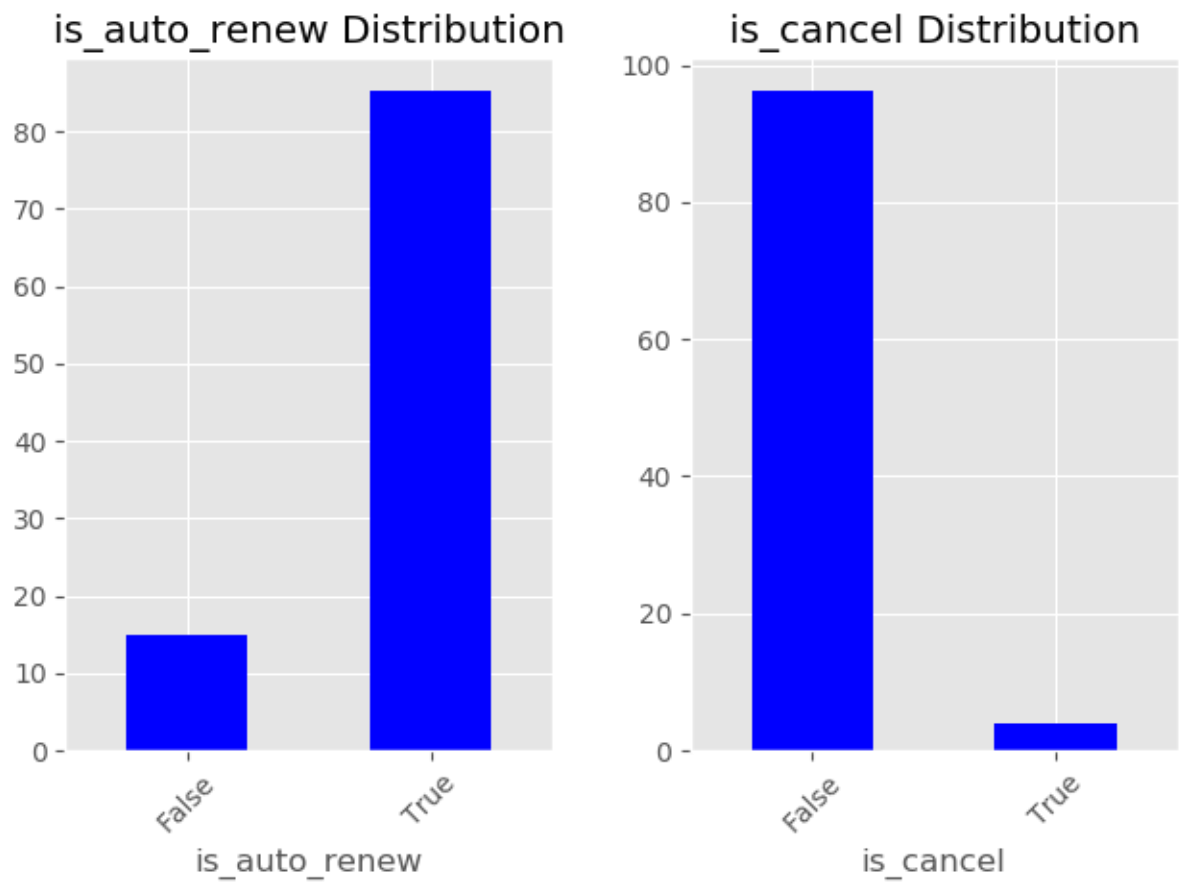
1) Transactions 1
2) Users log 8

1) Transactions

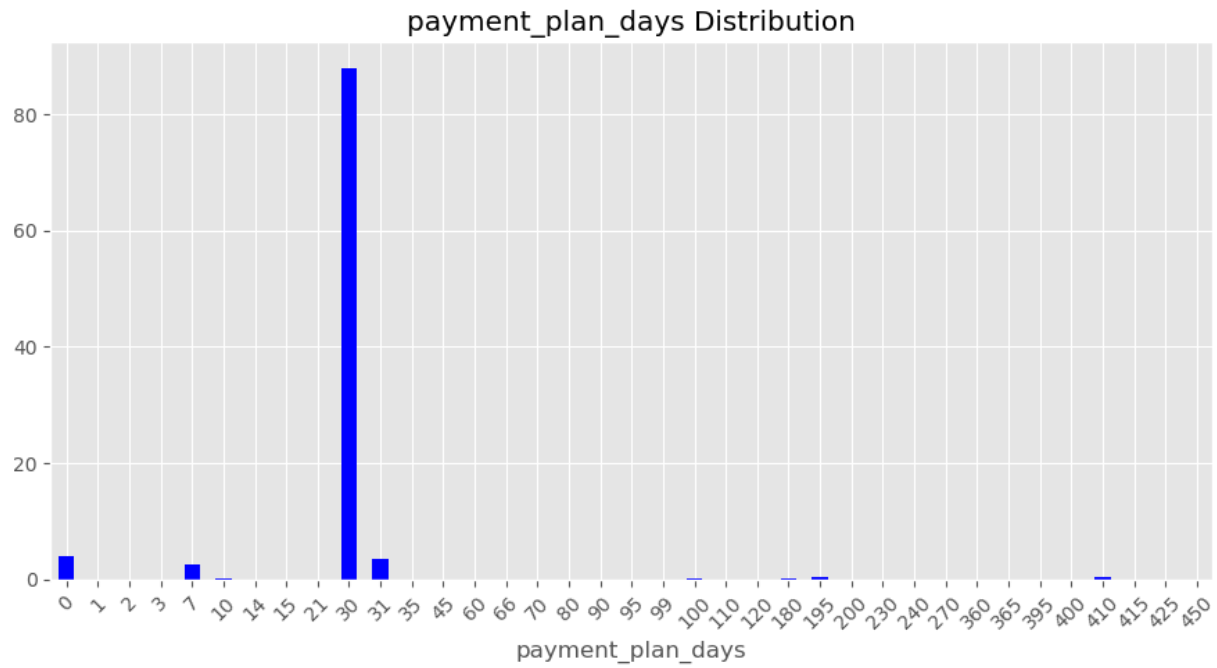
Based on roughly 21.5 Million transactions, more than 50% are completed using payment ID 41. Unfortunately, payment methods are represented by numbers only, so we don't know what it means. Intuitively, this may be representative of credit card automatic payment.



In this 2 years and 2 months period, membership automatic renewal is authorized 85% of the time. Looking at KKBOX website and using google translate (it's in Mandarin), there is a discount when automatic renewal is active which can explain its popularity. Users are allowed to cancel at any time, this is recorded as a Boolean. An overwhelming majority of transactions do not show a cancellation (~95%). This is consistent with the fact that churn rate is very low (e.g. 6% in February 2017). Also, it happens that a user chooses to cancel its current plan to join another one.



Plan's duration in days shows a wide variety compared to KKBOX website where 30 days, 90 days, 180 days and 365 days are listed. However, KKBOX does offer bonus days which adds multiple combination and their special offers may have changed over this 2 years period. The most popular plan by far is 30-31 days followed by 7 days. It extends up to 450 days. Note this data has 0-day plans which seems erroneous, it represents 4% of all transactions. Taking a closer look at 0-day plans, all of them are also missing list prices which have a value of 0 (We will worry about list prices later).



To replace 0-days payment plan, we need to consider multiple scenarios:

- User renews on time
- User renews late
- User is actively cancelling his/her plan

When a user renews on time, it means there is an overlap between the current transaction date and its prior expiration date. We will take advantage of this fact and thus use the time difference in days between successive expiration dates.

When a user waited past his/her membership expiration date then this is an independent transaction. Hence, current transaction date and expiration date time difference should tell us the plan duration.

Finally, when a user decides to cancel his\her membership, one can use the most recent plan he/she subscribed to before leaving.

At this stage, one may noticed some estimated plan duration were negative! It turns out some user's transaction history did have expiration date which went back in time. This behavior cannot be explained, it is most likely an error during data collection. Those events were discarded.

Furthermore, when there is a missing value for payment plan days on a user's first transaction, we do not have any prior history to estimate its plan duration. Users having a unique transaction leaves us clueless so we are removing them. For multiple transactions history, when both first and second transactions are automatic renewal then it must have

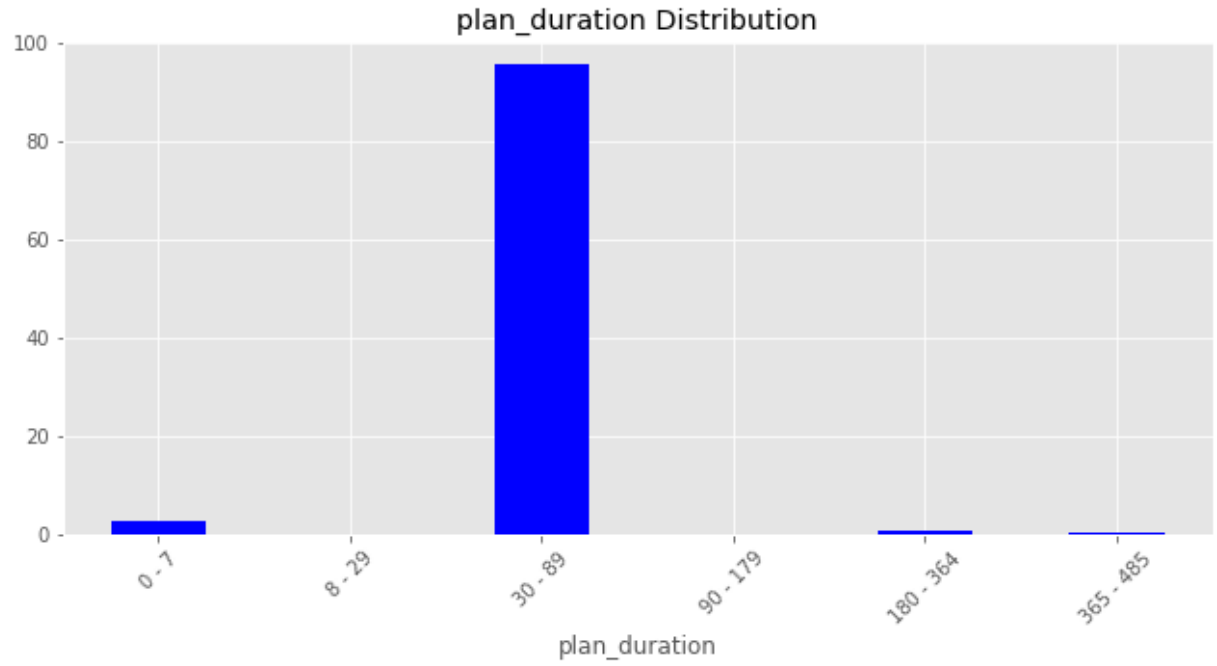
the same payment plan days. Except the latter scenario, there isn't any other case where we can be sure of the right payment plans days. Therefore, we wil discard all other user's first transactions (total of ~95k). Note that those transactions, including unique transaction, are still useful to determiner whether a user has churned.

Our estimated plan duration can be adjusted for monthly subscription. Because 30 days plans actually means monthly membership, our estimation varies between 28 to 32 days depending on the number of days in each consecutives month. Plan duration have been corrected to take this fact into account. Even the raw data has 31 days plan where it really is 30 days.

At this point, we actually created more payment plan than we had in our data even after monthly membership adjustement. To reduce the number of plan, we can group them by intervals:

| 0 - 7 | 8 - 29 | 30 - 89 | 90 - 179 | 180 - 364 | 365 - 485 |

Those intervals are based on the information available on KKBOX website. For instance, there is a 30 days and 90 days plan hence we make an interval between 30 and 89 days. The assumption is you can only have extra days in your subscription.

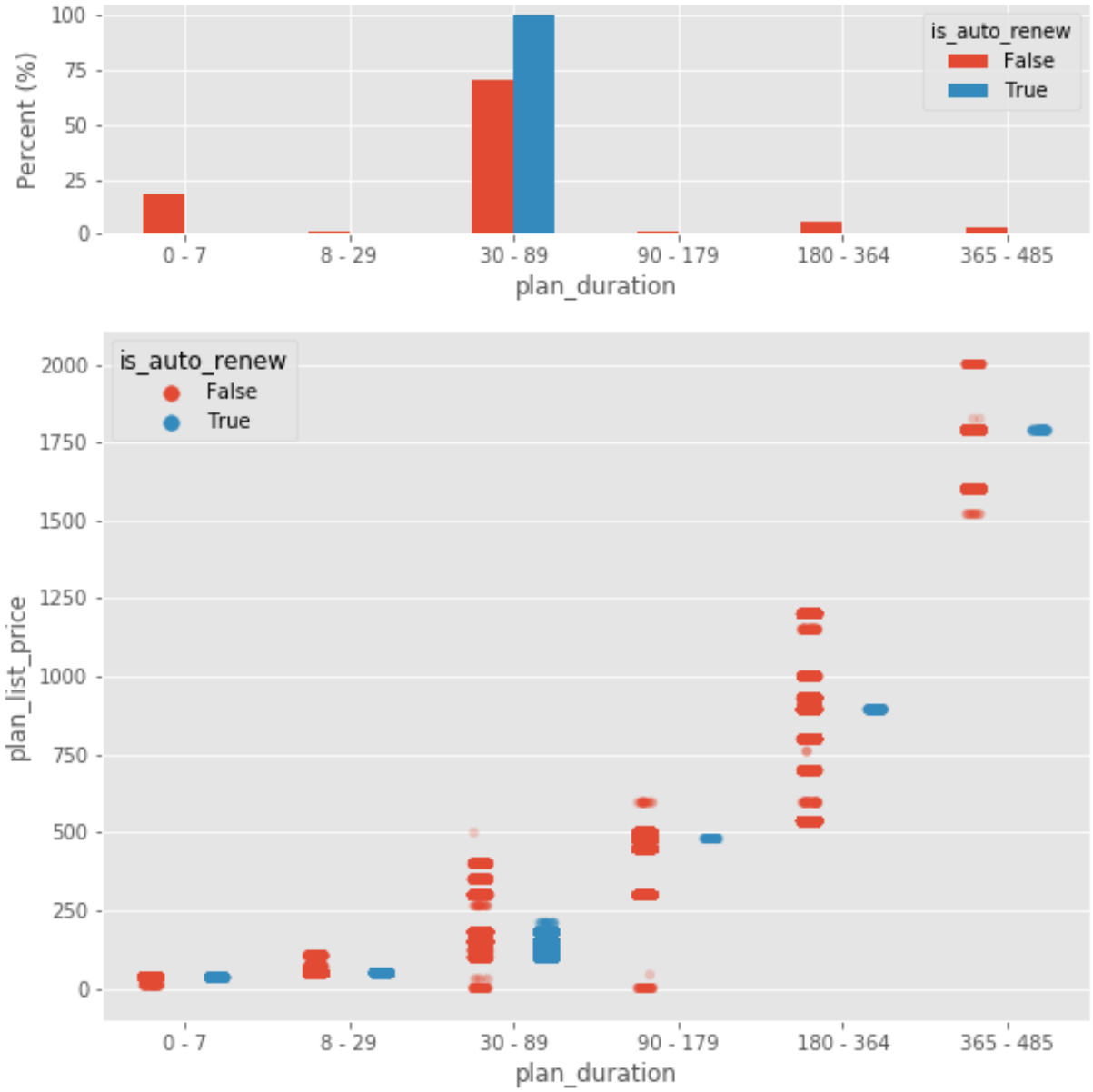


Plan list price remains below 150 NTD (New Taiwan Dollar) for 90% of all transactions. Most frequent prices are 149, 99 and 129 in descending order. There is a non-negligible number of 0 NTD list price (around 8%) which indicates missing values. Indeed, plan list price should always have an amount for each available plan regardless of how much

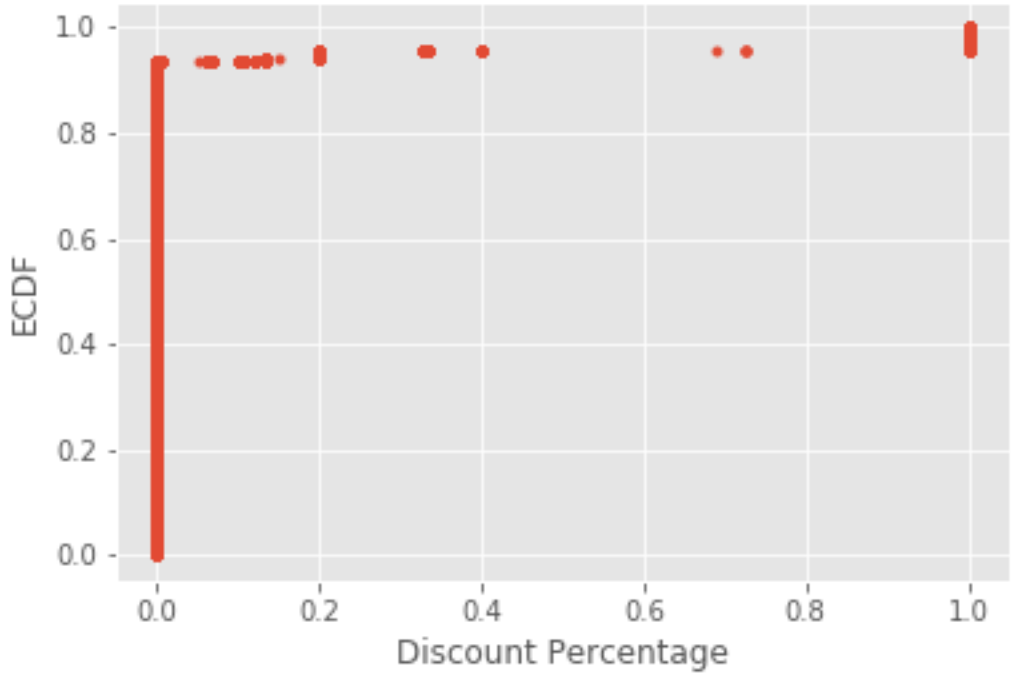
customers actually paid. We saw earlier that missing payment plan days also had missing list price. As a result of filling payment plan days, we have new plan duration. So we should find a mapping between plan duration and list price based on our data as opposed to payment plan days. Otherwise, we won't be able to map every plan duration to a list price. Each of the six plan duration intervals has been linked to the most significant list price as shown below:

Plan Duration in days	Plan list price in NTD
0 - 7	35
8 - 29	50
30 - 89	149
90 - 179	480
180 - 364	894
365 - 485	1788

All missing values have been replaced at this stage. Let's take a look at plan list price as a function of plan duration. As expected, longer plans tends to be more expensive. Interestingly, there are overlaps in list price except for long term plans (> 365 days) which are clearly more costly.

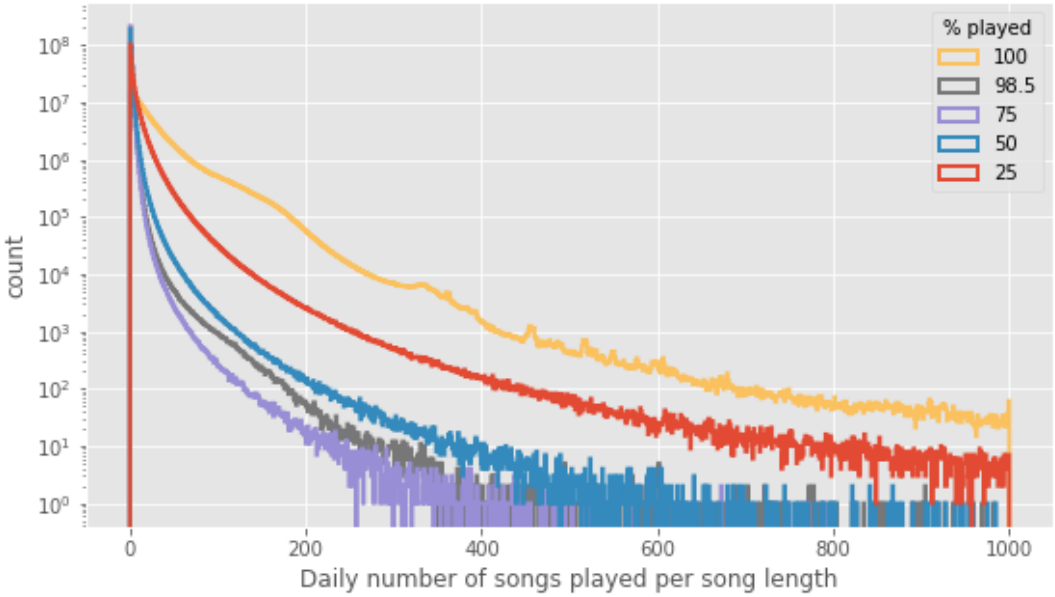


Finally, we will look into the difference between list price and actual amount paid. Is there any discount? Any trend? As a matter of fact, getting a discount is pretty rare. Less than 8% of transactions have discount! Surprisingly, there are negative discount! This means the actual amount paid is greater than list price. To correct for this anomaly, list price were matched to their actual amount counterpart.

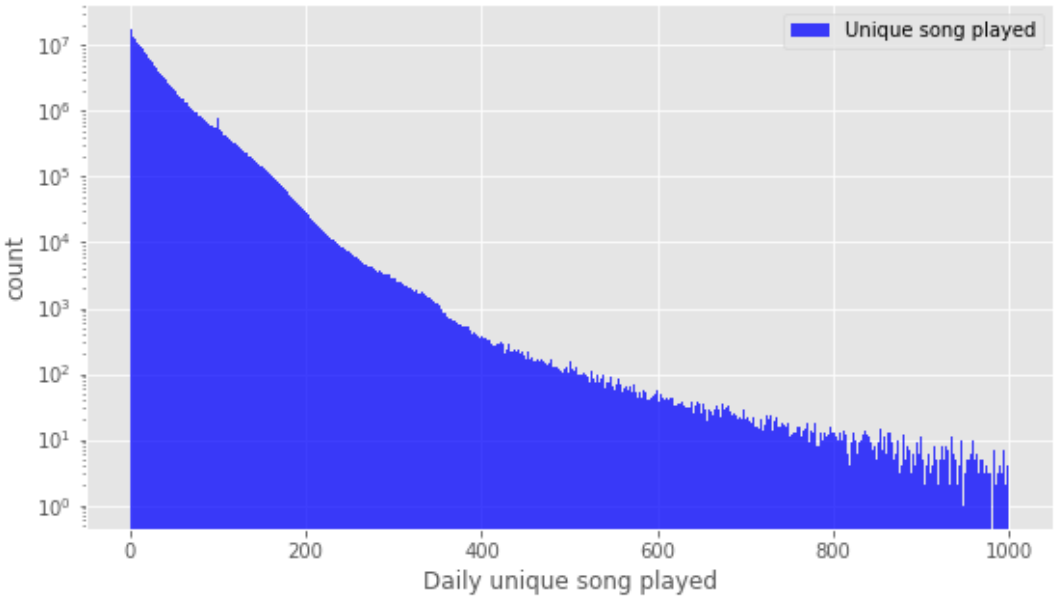


2) Users log

Users log records multiple metric related to user's listening habits. For instance the number of songs played per percentage of song length quickly decays as shown on the figure below. Percentage of song length are intervals. At 50% it represents number of songs played between 25% to 50% of song length. As the number of songs increase (>8 songs), there are more people listening to the whole song compared to 25% second followed by 50%, 98.5% and 75%.

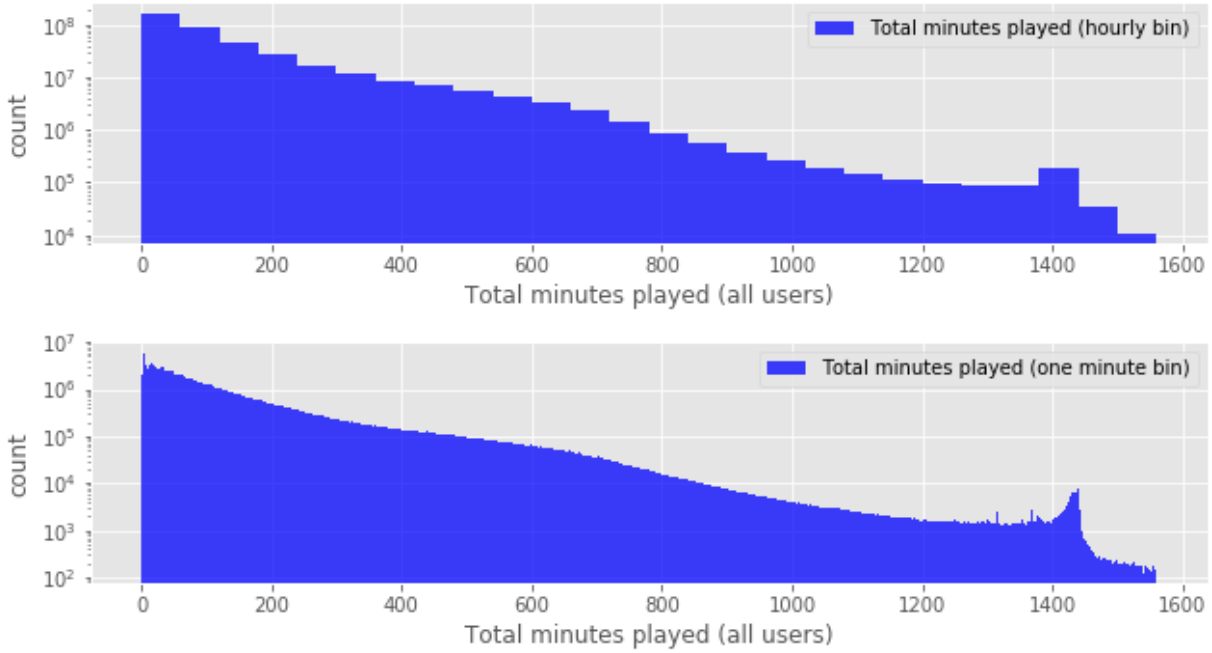


If we take a look at the number of unique songs played, it also decreases rapidly indicating most people don't have very long playlist.

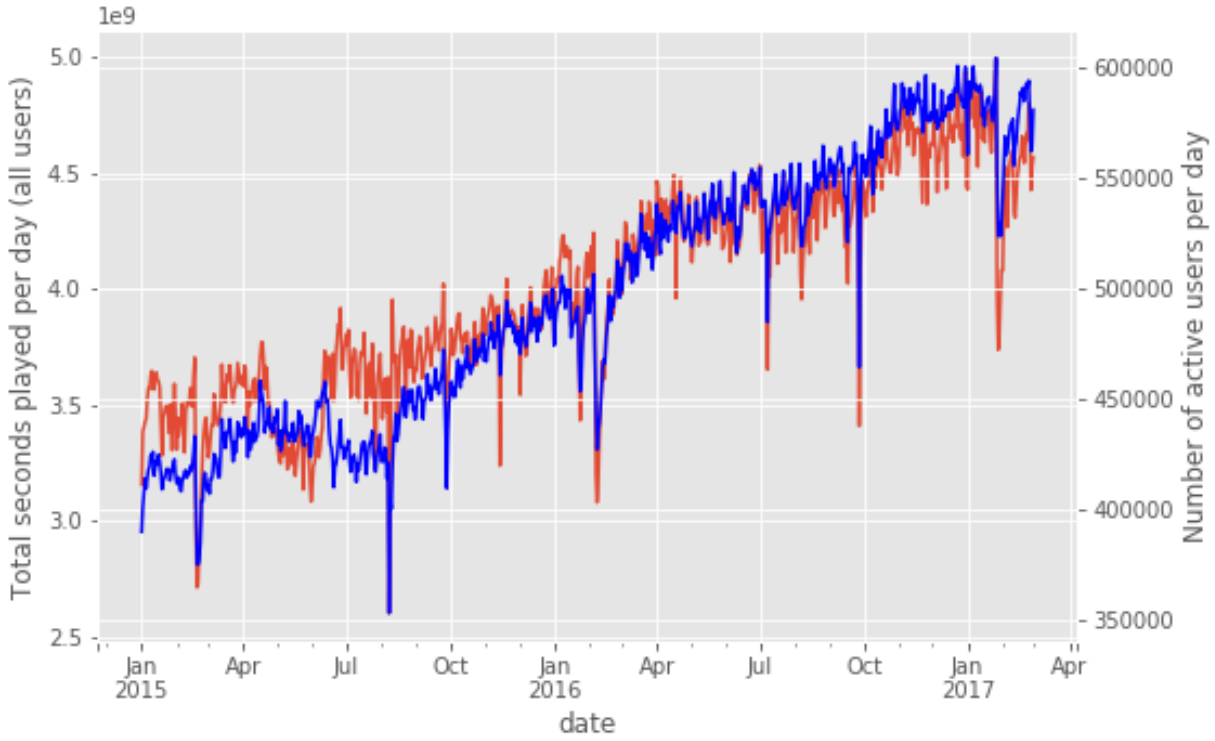


The total number of seconds played is consitent and declines as well. Surprisingly, there is a small peak (plots are log scale) at 24h and more data beynd 24h which should not happen

as it is a daily activity log. One scenario to explain this 24h peak would be 24/7 shops which relies on this streaming service for their background music. The most frequent time is 3-4 minutes of daily music. There are some peak forming every 15min which probably means some people have busy schedule and just listen to a playlist that accommodate their free time.



The last interesting plot is to look at the total listening time per day. Overlapping the number of active users each day, we can clearly that both total listening time and number fo active users correlate very well. Some values were extremely negative or positive so I filtered those out by making sure total listening time is between 0 and 24h.



There are multiple drops occurring in user's count and more so in listening time. Late January drops corresponds to Chinese New Year which last 5 days not including new year's eve.

Another periodic drop is on Father's day (always August 8th). Note that in 2015, father's day was on a Saturday while father's day in 2016 was on a Monday. Thus we can see a lesser drop on Sunday, August 7th in 2016

Decline on September 28th, 2015 corresponds to mid-autumn festival. It does repeat in 2016 with lesser effect (September 17th).

The closest event to the large drop on September 27th, 2016 was Teacher's day (birthday of Confucius) that took place on Sep. 28th, 2016.

There is an extended drop in May 2015. It could be due to technical malfunction (music app not working for iphone for instance).