After struggling for a while to reshape and filter the different datafiles and merge them into a sample that can be held in memory and perform viable dataframe operations upon I ended up with a sample of the available data from which I can infer some things.

The resultant dataframe contains features of the user id’s in the train set of which 20,618 are non churners and 1,199 are churners. I filtered out daily user log data for the last 30 days for these users including listening percentages of songs played, number of unique songs played and total listening time.

I did some quick data exploration comparing aggregated values for churners and non churners. I also used pandas crosstab function to show visualizations for cross tabulation of categorical variables for which an average value wouldn't mean much.

I used the pandas corr function which computes correlation coefficients for all the columns of a dataframe. You can quickly see some things that appear to negatively correlate with churn including age, number of songs played to 100%, number of unique songs, and seconds listened to. Something that appears to positively correlate with churn is the number of songs played to less than 100%.

To test for the significance of these correlations the stats module in SciPy can be used. The pearsonr function calculates both the Pearson correlation coefficient and its p-value from two arrays. In this this case the p-value is used as a rough estimate of the probability of an uncorrelated system producing a correlation coefficient whose magnitude is greater than or equal to this one.

The p-values for the number of songs played from 0-25%, 25-50%, and 75-98.5% all had values greater than 0.45. The correlation for songs played from 50-72% had a p-value of 0.06. Age, songs played to 100%, and total seconds had p-values 0.001 or lower. The features whose p-values were lower than 0.1 have shown with roughly 90% confidence that they are correlation with user churn and should be valuable in constructing a prediction model.