Hendon Mob Analysis

Brendon Kaufman

2/5/2019

Hendon Mob Database Analysis - IN PROGRESS, UNFINISHED

Background: The game of poker is arguably America's most popular card game. Based on the variant, players hold various numbers of cards, trying to make the strongest 5-card combination. The strengths of these combinations are determined by a ranking system which is uniform across all variants. Poker's biggest attractions are its unique mix of luck and skill and the fact that it's one of few games where players regularly bet money as part of the game. These factors draw a wide variety of players, from professionals who make their living solely from playing the game, to recreational players who enjoy a challenge or a gamble. Although there are clear differences between these populations, there has yet to be rigorous analysis using existing databases which would demonstrate how these differences manifest themselves.

Goal of this analysis: I propose an analysis of poker's only public database of player performance, The Hendon Mob tournament database, to reveal insights about different populations playing tournament poker.

Data : The data analyzed come from The Hendon Mob, a tournament poker database which displays all live cashes for individual players. If a player made money in an official tournament, it is recorded on www.thehendonmob.com. I used the package `rvest` to scrape data from individual player pages. Then, I created functions to extract and summarize the most important statistics for each player and create a summary dataframe, where each row contains one player and his defining statistics and information. This is done with the functions in the script *01_scrape_hendon_mob* located in my Github. The script *02_analyze_hendon_mob* allows the user to convert the summary dataframe into a format suitable for analysis, and provides some sample analyses.

Analysis of entire Hendon Mob population

The Hendon Mob touts itself as the "world's largest live poker database," containing information on **579,387** players as of February 6th, 2019. It is possible to scrape the entire database, but this would take a long time, especially when we need to adjust the scraping to add or remove elements. Therefore, the script *01_scrape_hendon_mob* allows for the user to choose how many players they would like to scrape statistics for, then output this into the aforemntioned summary dataframe. The function does this by randomly generating player urls, and, should they be valid urls, scraping the information found at that url.

Therefore, the summary dataframe contains **randomly selected** players from the Hendon Mob, so that we can approimate what the population looks like without downloading the entire database. At the moment, I have downloaded **5,439** players into the *hendon_summaries* csv located in the repository, or about 1% of the website. I attempted 6,000 randomly selected players, but about 600 of the randomly created urls were invalid. This sample should reasonably approximate the behavior of the entire database. Below is the first row from the sampled database.

```
hendon_summaries_df[1,]
```

```
## # A tibble: 1 x 16
##   name  nationality average_buy_in number_of_cashes sum_of_cashes
##   <chr> <chr>               <dbl>            <dbl>         <dbl>
## 1 Davi~ Belgium              7054              134       9879441
## # ... with 11 more variables: average_cash <dbl>, average_placement <dbl>,
```

```
## #   number_of_binks <dbl>, binks_proportion <dbl>,
## #   number_of_countries_cashed <dbl>, first_date <date>, last_date <date>,
## #   years_played <time>, average_time_btwn_cash <dbl>, unique_views <dbl>,
## #   quantile <int>
```

Fields

Some fields in the summary dataframe are scraped directly, others are modified using post-hoc manipulation. The list of all of the fields which are in the summary dataframe are is the following.

1. **name** - Playér's name
2. **nationality** - Player's nationality
3. **average_buy_in** - Player's average buy-in, in USD

i) Note that this item is imperfect because buy-in is not always listed, and when it is, sometimes the currency is difficult to guess. I assume here that the currency of the buy-in is in the currency of the country of the tournament, however this is not always true (and there is currently no better way). For example, some events in Ukraine transacted with USD, others with Ukrainian Hryvnia. Since I automatically convert all foreign currency to USD based on 2017 exchange rates, this results in some values being converted which actually did not need to be, and therefore in the buy-in values being wrong.

4. **number_of_cashes** - Number of events cashed in career
5. **sum_of_cashes** - Total amount of money cashed for in poker career, in USD
6. **average_cash** - Average amount cashed for per tournament
7. **average_placement** - Average placement in tournaments
8. **number_of_binks** - A bink is poker slang for a sizeable tournament poker score. There's no agreed upon definition of a bink, but I define it here as any cash above 20 times the average buy-in. This field counts all of those cashes. For the reasons noted above, this field is somewhat unreliable since it depends on average buy-in.
9. **binks_proportion** - The percentage of tournament cashes that were binks
10. **number_of_countries_cashed** - The number of distinct countries that a player had a tournament cash in.
11. **first_date** - The date of the earliest tournament cash for a player.
12. **last_date** - The date of the most recent tournament cash for a player.
13. **years_played** - The difference between the date of a player's most recent cash and their first cash, in years. Note that this does not assume that the player has continued to play since their last cash.
14. **average_time_btwn_cash** - The average number of days between a player's cashes, determined by using the first and last cashes as the endpoints
15. **unique_views** - Number of unique views of the player's profile
16. **quantile** - Players are separated into 4 quartiles based on their total cashes. Players in quartile 1 are the 25% of players with the least amount of cashes, while players in quartile 4 are the 25% of players with the most amount of cashes.

To get a feel for what our data look like, I'm going to perform some basic calculations on the entire database.

Average earnings

First, we are interested in the average earnings of a player in the database

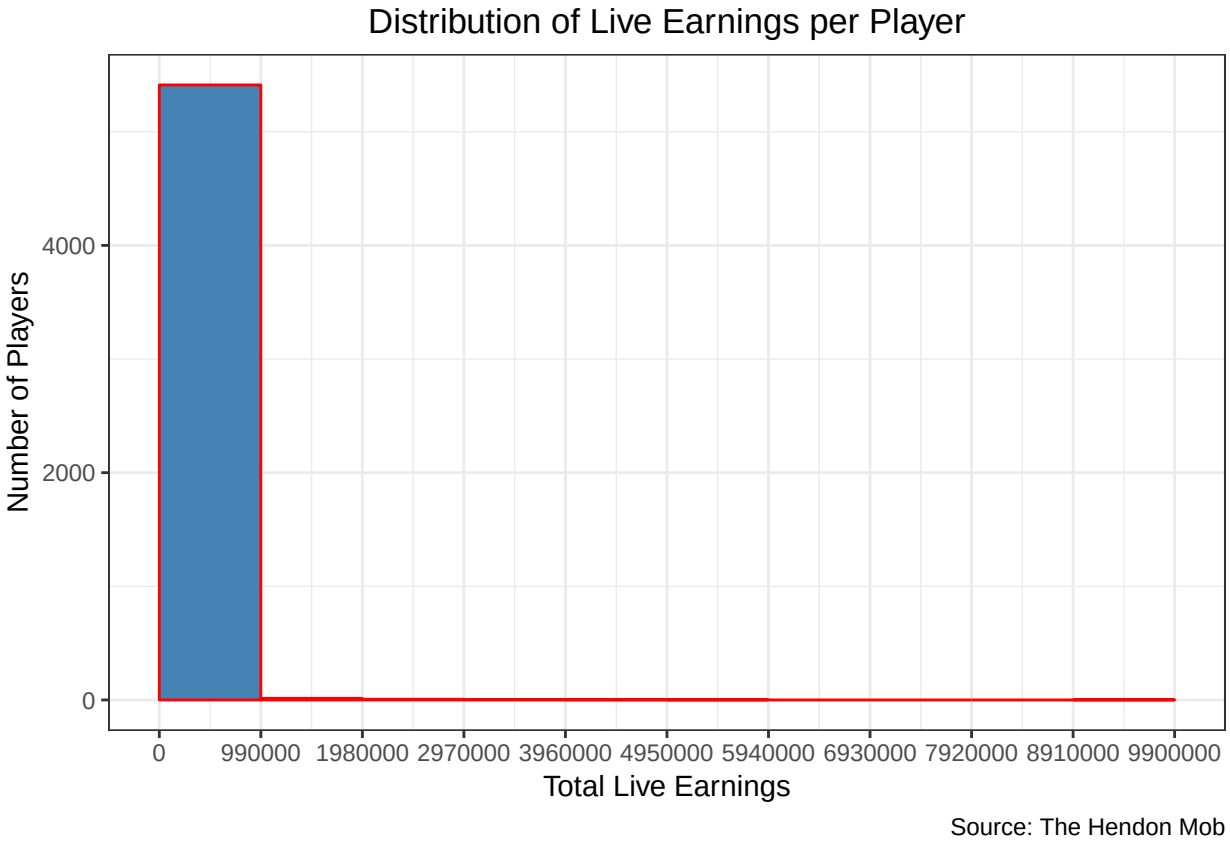
```
mean(hendon_summaries_df$sum_of_cashes)
```

```
## [1] 31835.45
```

The average earnings of a player in the database is **$31,835.45**

Distribution of earnings

How are these earnings distributed among the players? Poker tournaments are competitions who award large sums of money to small amounts of the field, and little to nothing to the rest. We would expect the distribution of earnings to follow accordingly, likely being heavily positively skewed with few players earning large amounts of money and the rest earning very little.

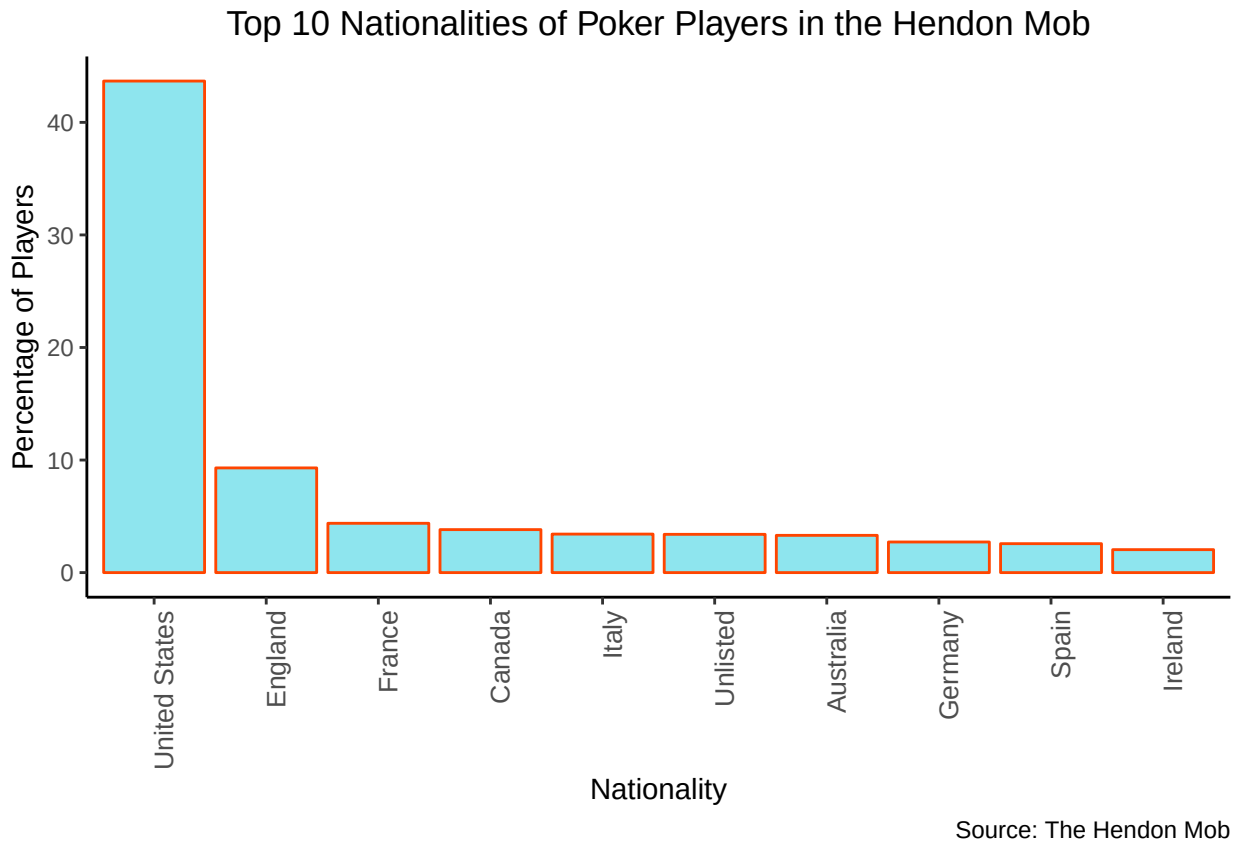


We find that the highest earner in our sample is Davidi Kitai, earning **$9,900,000**. However, almost all of the players in the sample have earned less than 10% of his earnings. The earnings distribution is therefore heavily positively skewed, and would likely hold that distribution even if we were to zoom into the population of lower earners. We will do this later in the analysis.

Distribution of nationalities in the sample

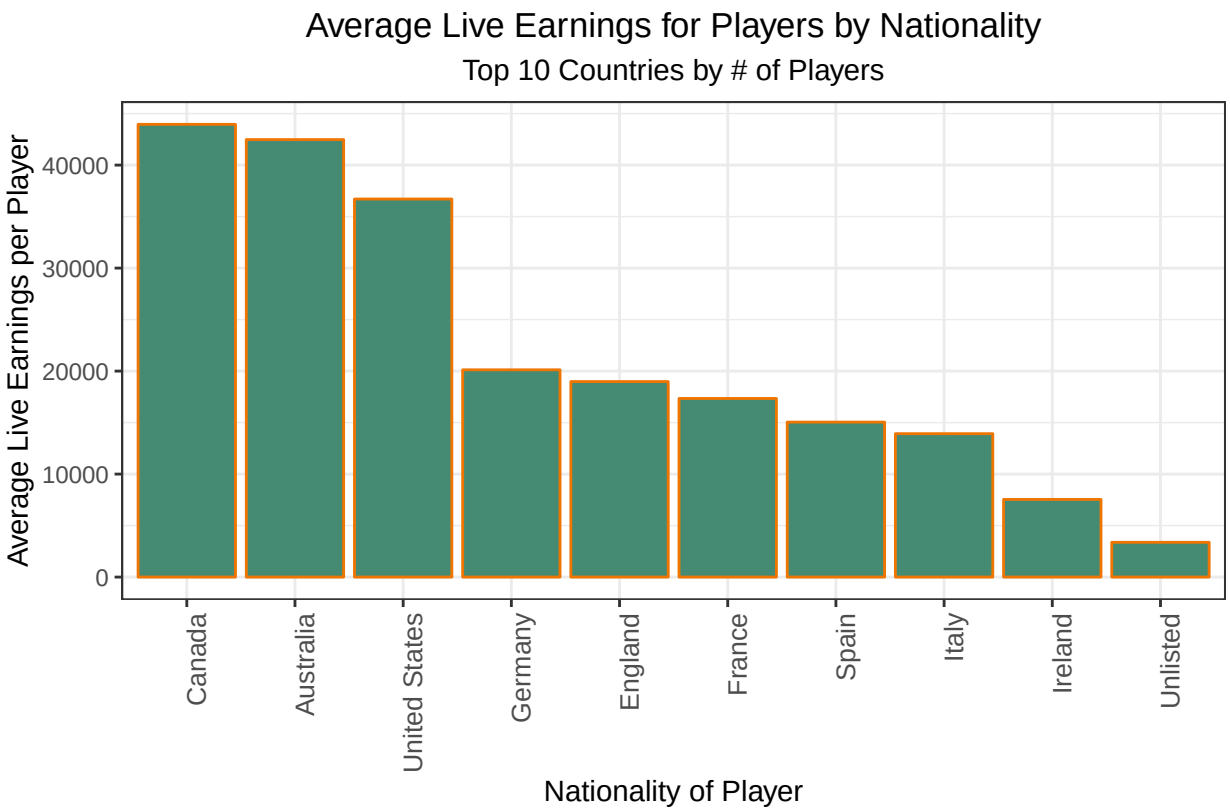
We might be curious about the nationalities of players in the Hendon Mob database. Poker is obviously popular in the USA, but it is also hugely popular in parts of Europe and Asia as well. How does this manifest

itself in the sample?



We find that Americans dominate the database, with over 40% of the total players being from the US. The English come in second with 10%, followed by others such as France, Canada, Italy, Australia, and Germany.

Average earnings per country



In this sample, Canadians have the highest average earnings, followed by Australia. However, because of the highly skewed positive distribution, it is likely that these numbers are highly influenced by the biggest earners in the sample. This sample drew many high-earning Canadians, Australians, and Americans, but few high-earning Europeans. It would be interesting to see what other samples would return, but since this is a rather large sample, we can estimate that this distribution is approaching the population distribution.

Average number of years played in the sample

How long have most of the players played in the sample? Note that with this data we can only find the time between the first and most recent cashes, or the time between the first cash and today. The latter assumes that all players have continued to play poker until today and the former assumes that players have not played after their most recent cash. It is quite easy to go a long time without cashing a tournament, but here we define total time played as time between first and most recent cashes.

Since we cannot calculate years played for players who only have one cash, they are excluded from this analysis.

```
hendon_summaries_df_more_one_cash <- hendon_summaries_df %>%
  filter(number_of_cashes > 1)

mean(hendon_summaries_df_more_one_cash$years_played)
```

```
## Time difference of 4.341405 days
```

The average time played per player in the database (with > 1 cashes) is 4.34 years.