Hendon Mob Analysis

Brendon Kaufman

2/5/2019

Hendon Mob Database Analysis

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 appoimate what the population looks like without downloading the entire database. At the moment, I have downloaded **5,473** 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 Blai~ United Sta~           2933               77       4359156
## # ... 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

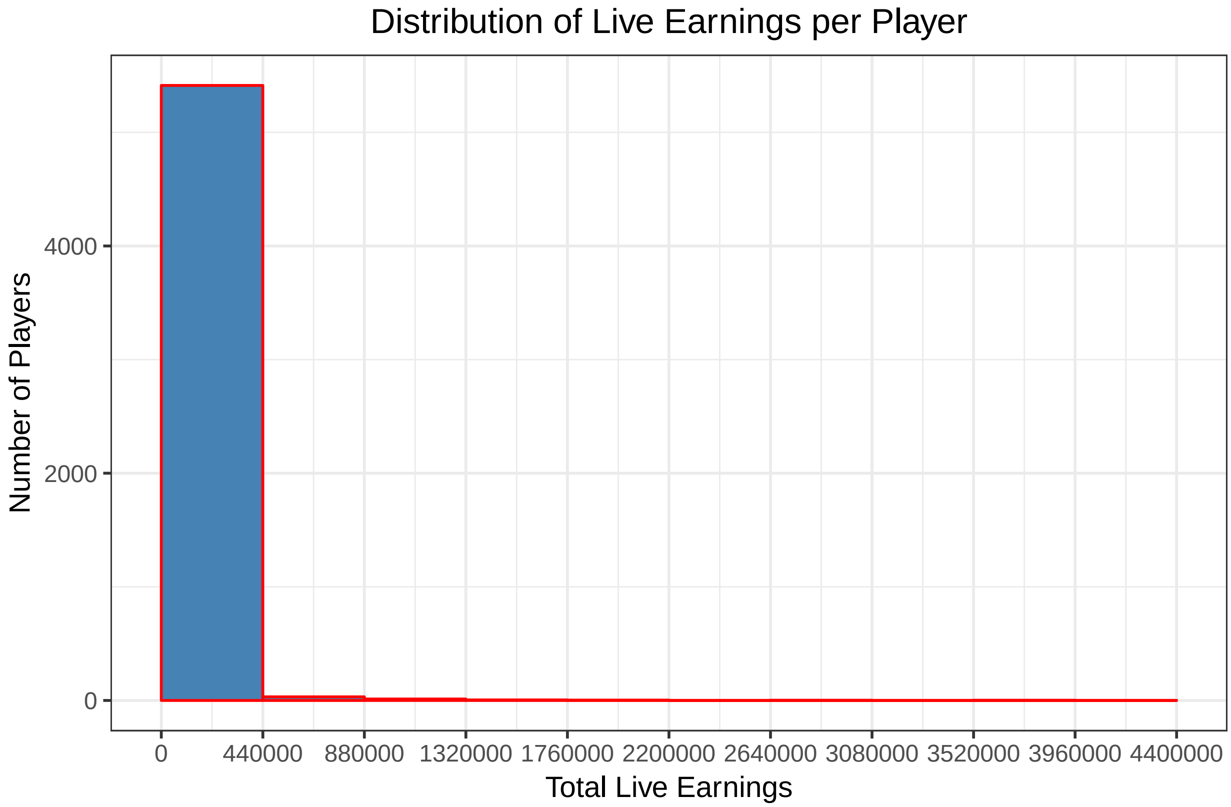
```
mean(hendon_summaries_df$sum_of_cashes)
```

```
## [1] 27863.98
```

The average earnings of a player in the database sample is **$27,863.98**

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.

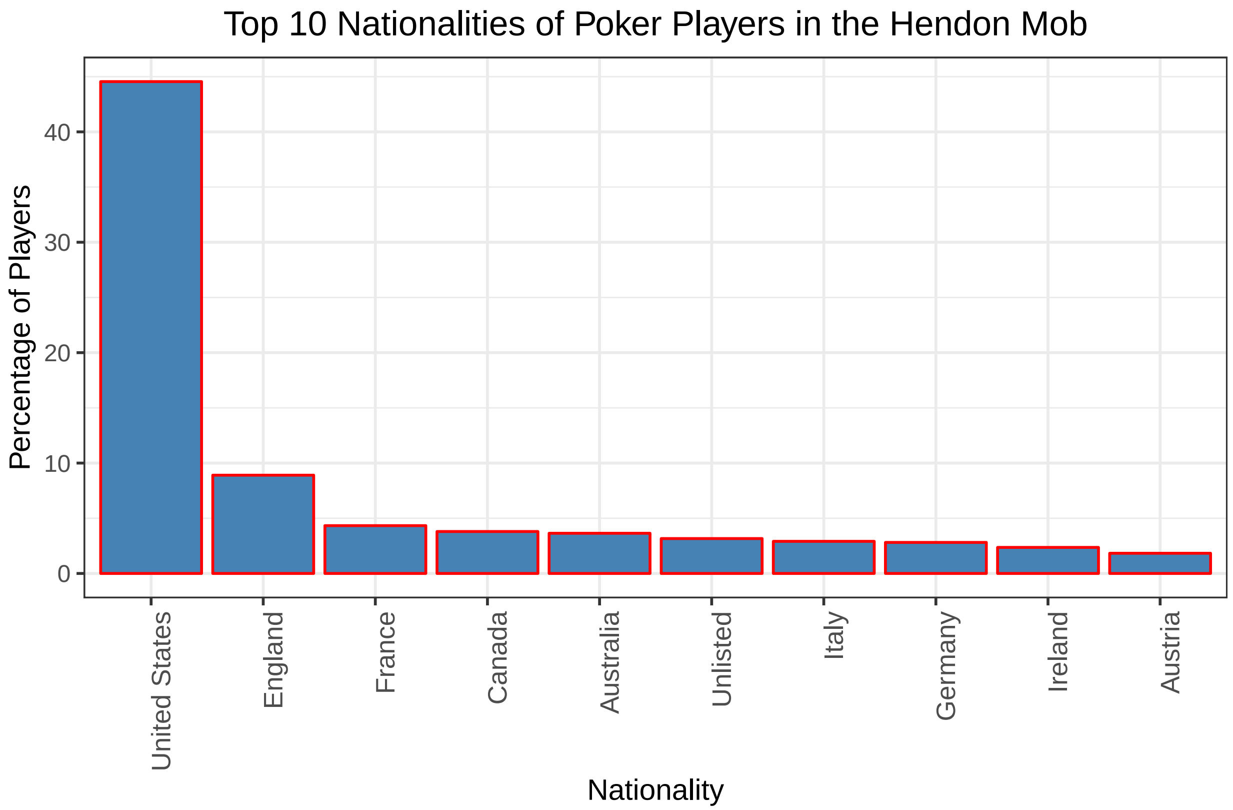


Source: The Hendon Mob

We find that the highest earner in our sample is Blair Hinkle, earning **$4,400,000**. 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 we would likely see the same 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

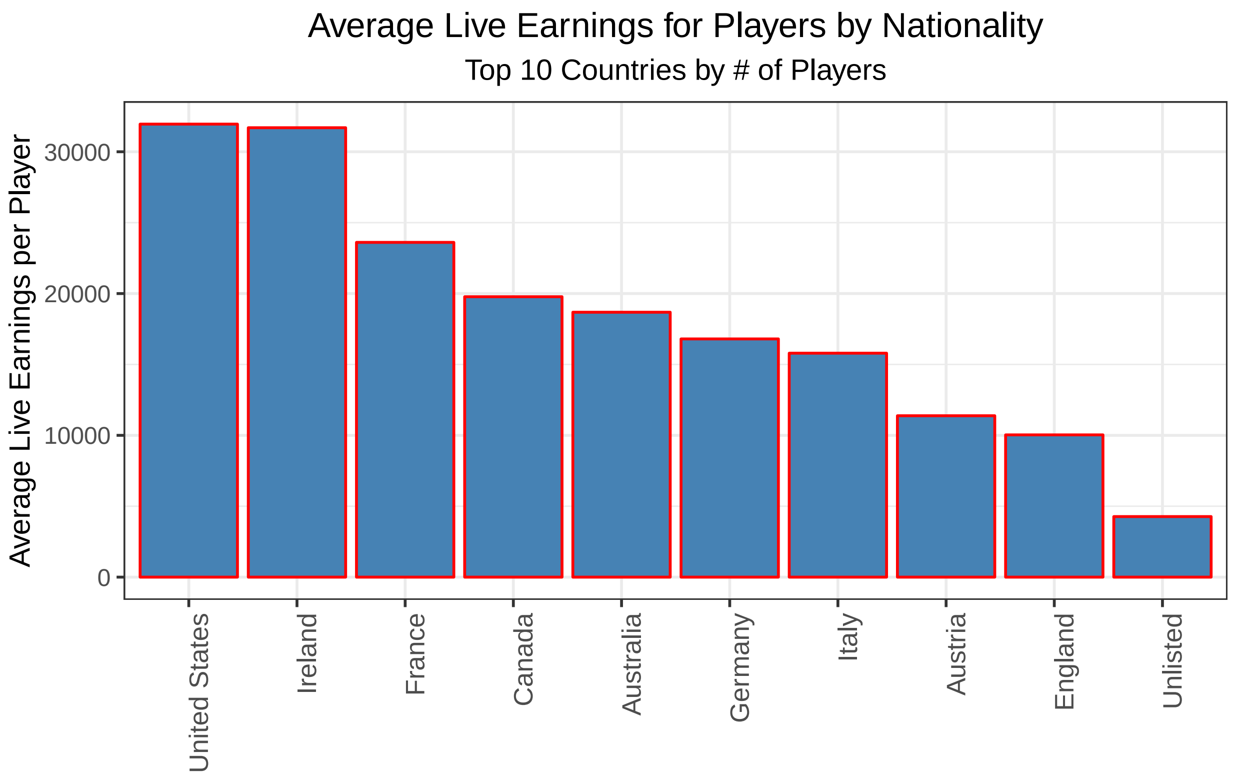
We might be curious about the nationalities of players in the Hendon Mob database. Poker is obviously popular in the USA, but it is also hugely popular in parts of Europe and Asia as well. How does this manifest itself in the sample?



Source: The Hendon Mob

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



Source: The Hendon Mob

In this sample, Americans 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 Americans, Irish, and French, but few high-earning players from other countries. 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 4.383461 days
```

The average time played per player in the database (with > 1 cashes) is 4.38 years.

Average number of cashes

Here we simply find how many cashes an average player has in the sample.

```r
mean(hendon_summaries_df$number_of_cashes)
```

```
## [1] 5.480724
```

The average player has 5.48 cashes in our sample of the Hendon Mob database.

Average buy-in?/Average buy-in distribution

Even though there are the aforementioned issues with the average buy-in calculations, we will take a shot at calculating the average buy-in of the players in the Hendon Mob database sample. We will exclude players who have an average buy-in of 0, as this clearly indicates the buy-in was miscalculated. I will also exclude players with an average buy-in between 0 and 10 dollars, which usually indicates a miscalculation, even though there are some very small buy-in tournaments, especially in casinos in the UK.

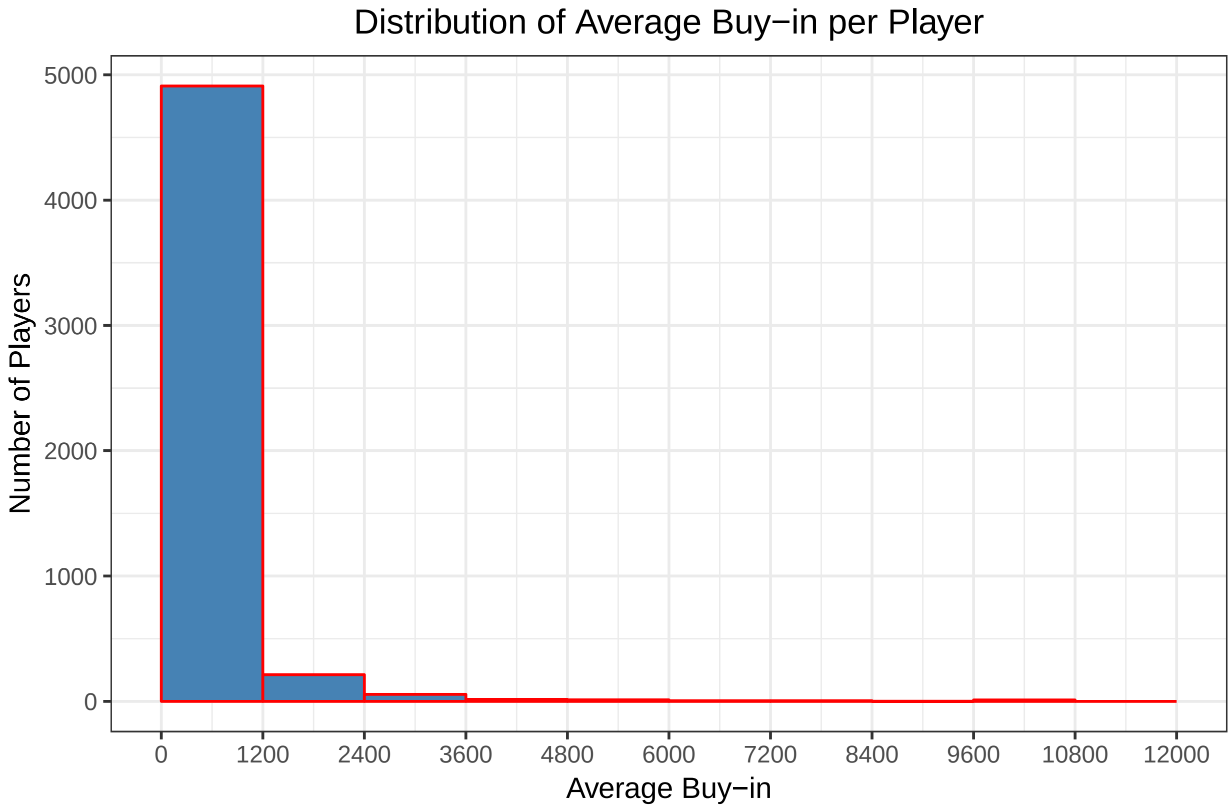
```r
hendon_summaries_for_average_bi <- hendon_summaries_df %>%
  filter(average_buy_in > 10)

mean(hendon_summaries_for_average_bi$average_buy_in)
```

```
## [1] 410.6351
```

The average player has an average buy-in of about 400 dollars.

Continuing the analysis of average buy-in, we are interested in how these average buy-ins are distributed in the population. Intuitively, the buy-ins should be distributed similarly to the cashes, in that most players buy-in for small amounts, but there is a population of high rollers as well.



Source: The Hendon Mob

The distribution of average buy-ins is as expected. The average buy-in levels range from 0 to 12000, with the average buy-in level around $400. Very few players have an average buy-in level greater than 3600.

Summary of analysis of total database

Our analyses aimed to give a general picture of the Hendon Mob database. To summarise we found that:

- Average earnings across all players: **$27,863.98**
- Maximum earnings in this sample: **$4.4 million** (Blair Hinkle)
- Earnings are incredibly positively skewed, with more than **90%** of the players earnings less than **$440,000**
- Most players are in the sample are from the **US** (40%) followed by **England** (10%), **Canada**, **France**, and **Italy**.
- **Americans** have the highest per player earnings in this sample at **$30,000** followed closely by **Ireland**. This number would likely change in another sample.
- The average time between first and last cashes is **4.38 years**.
- Players have **5.48** cashes on average
- The mean buy-in amount for players is about **$400**
- The distribution of average buy-in levels of players is positively skewed. The maximum average buy-in is **$12,000**, but less than **10%** of players have average buy-ins above **$3000**.

Analysis by Total Cashes Quartile

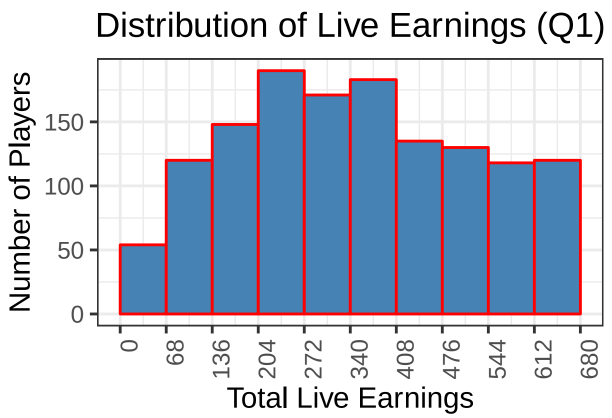
Now that we have an idea of what's going on generally in the database, we can start to look at differences between categories of players. One of the most salient differences between tournament poker players is how much they've earned in their career. It must be said that Hendon Mob earnings aren't a good indicator of how much players have *profitted* in their career, since losses are not recorded by The Hendon Mob.

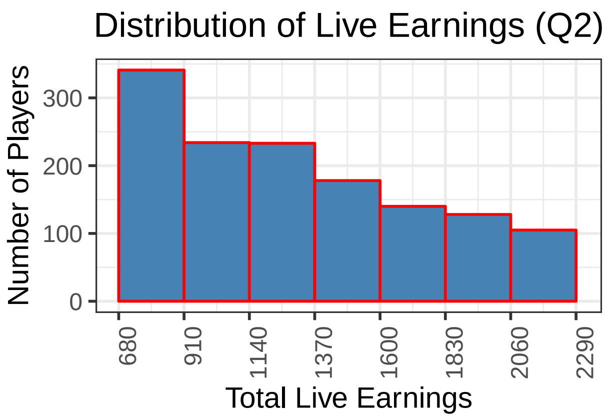
Therefore, if separate players into categories based on their winnings, we might understand what drives a player's total earnings. Have players who have cashed for more played for longer? Do they cash tournaments more frequently? These are the types of questions I intend to answer in this section.

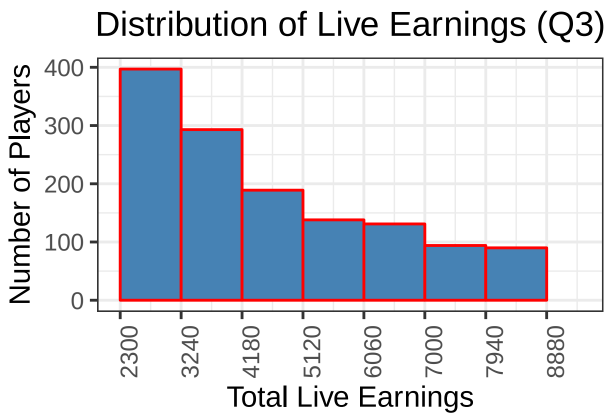
In this section I will be referring to players by **quartile** where players in quartile 1 have the least amount of cashes and players in quartile 4 have the highest amount of cashes.

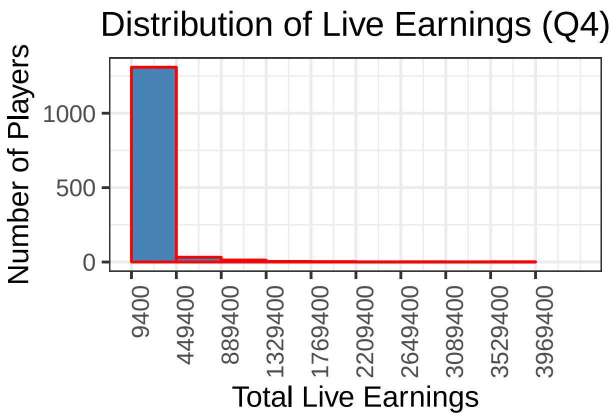
Distribution of total earnings

In this section, we try to get a basic understanding of the data by comparing the distributions of total earnings among players in the four quartiles.





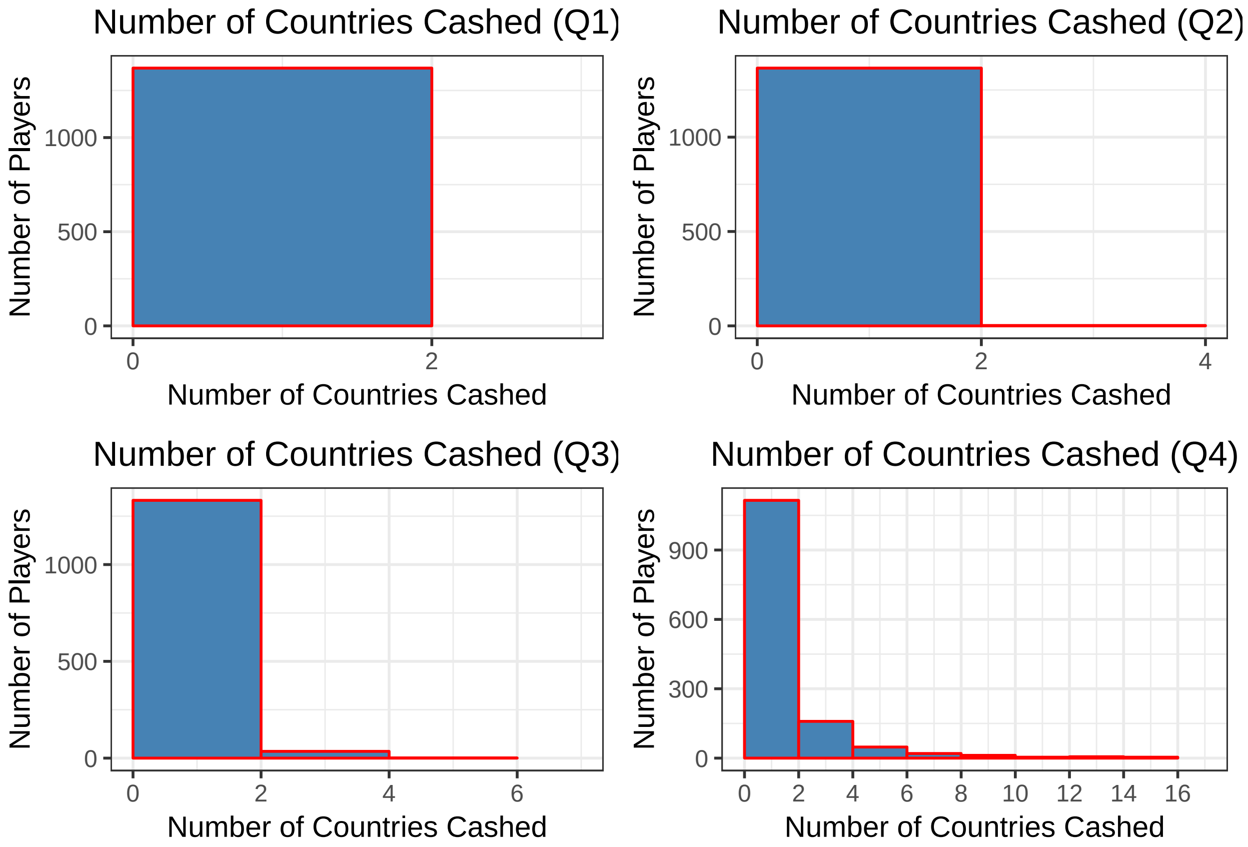




If we compare the four quartiles, we find that live earnings are more positively skewed quartiles of players with higher players. This means that among most of the population, earnings are somewhat uniform. However, once we go into the top 25%, we find that earnings are incredibly top-heavy.

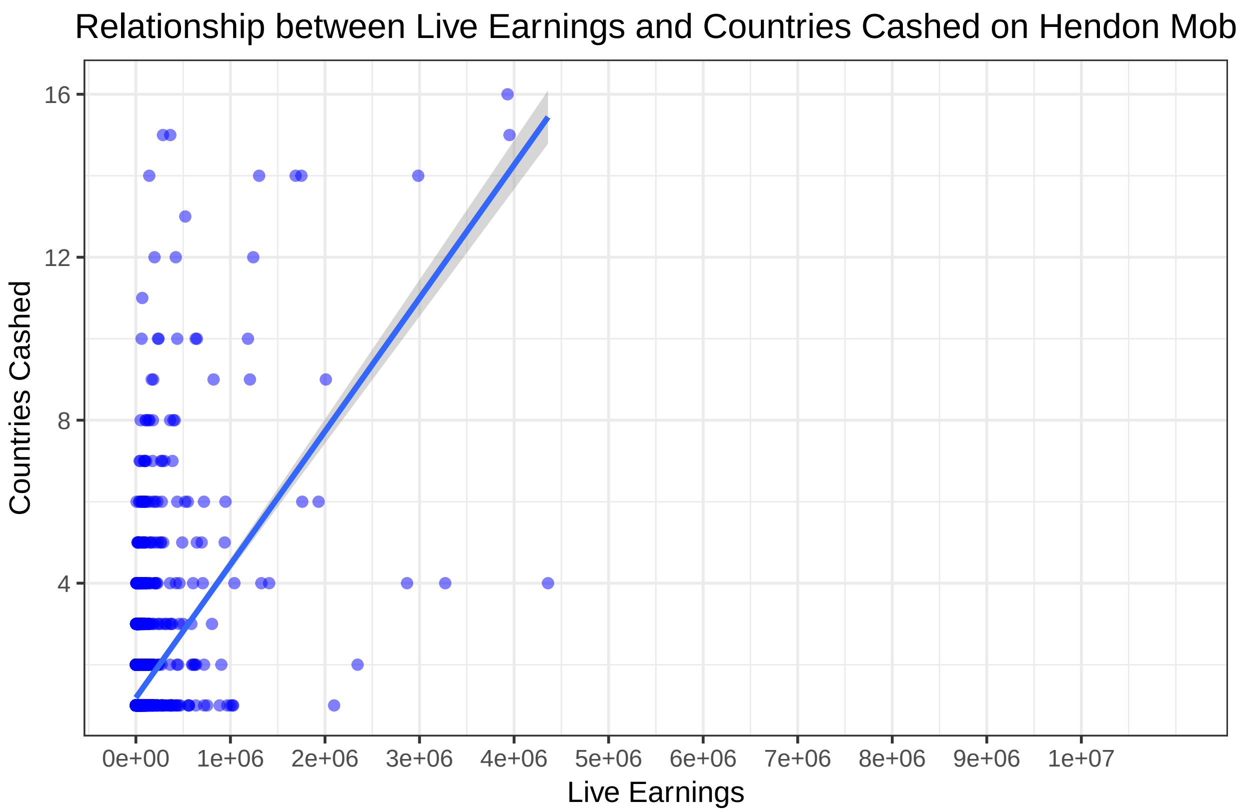
Distribution of Countries Cashed in

Here I compare the amount of countries that players have cashed in among the four quartiles. Do players who have earned more travel significantly more?



Players who have cashed for more money have had success in more countries, but still the vast majority of players have cashed in a maximum of 2 countries.

To visualize the positive relationship between earnings and countries cashed, I use the scatterplot below.

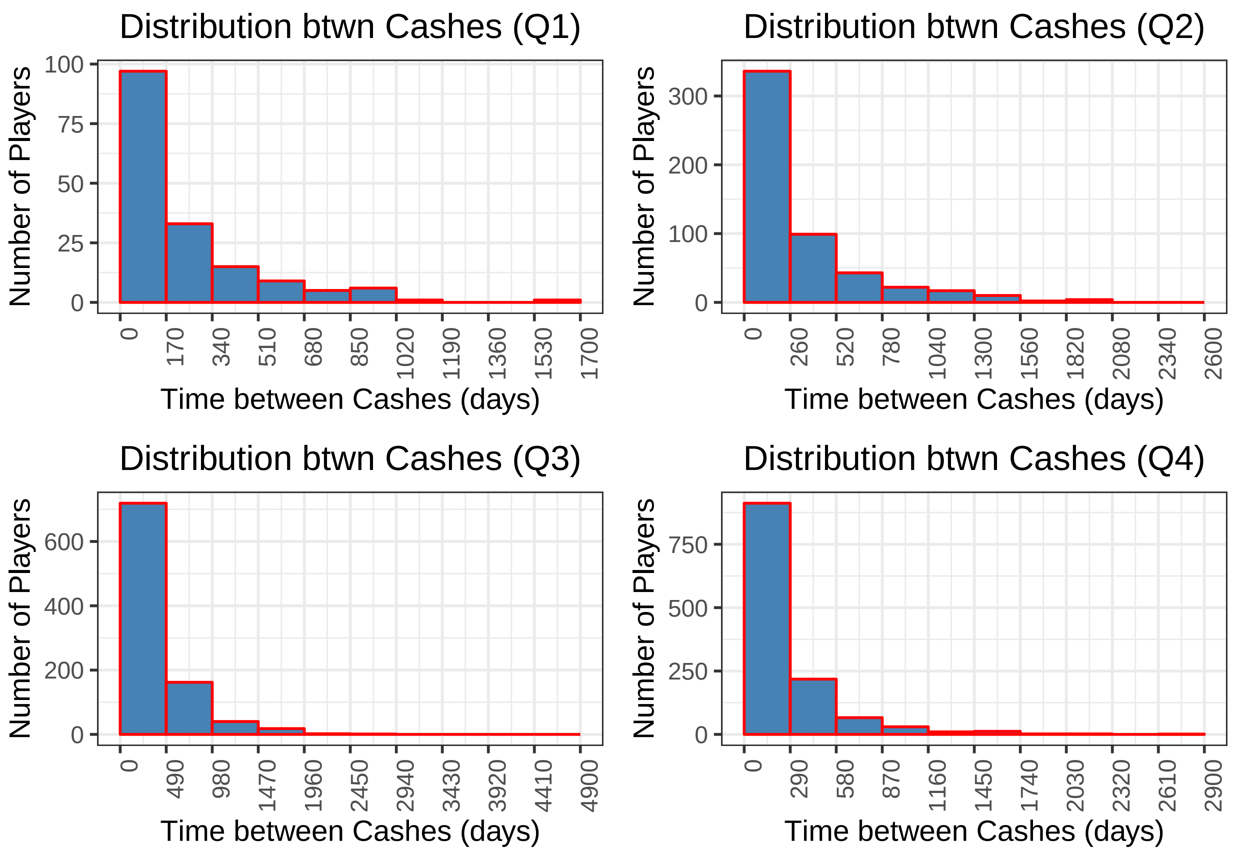


Source: The Hendon Mob

Indeed, there is a positive relationship between live earnings and the amount of countries a player has cashed in.

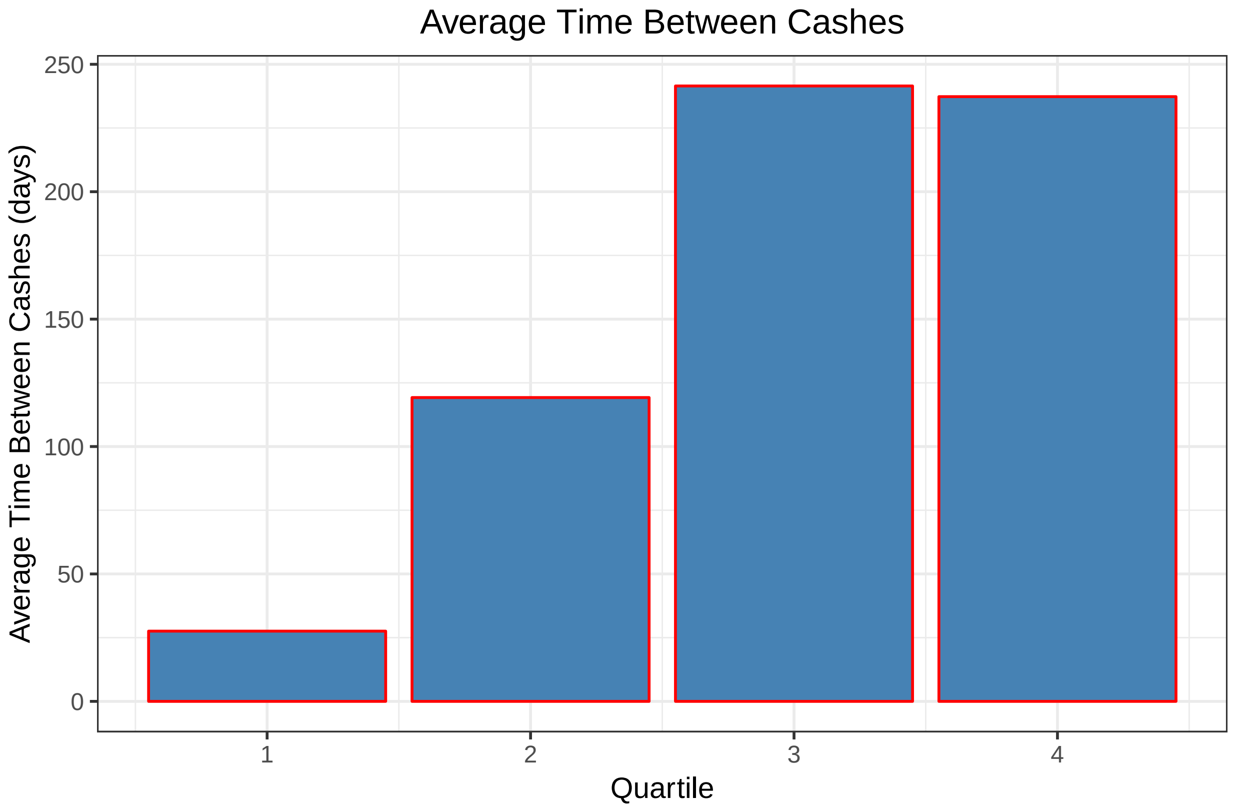
Distribution of time between cashes

In this section, I compare frequency of play among quartiles. To do this, I make use of the *average time between cashes* field. Players that play more frequently will have a lower average time between cashes. Intuitively, it would make sense that players with higher amounts of cashes would play much more frequently, but is this the case?



All distributions across quartiles are positively skewed, indicating that most players are grouped together on the lower end of the distribution, but very few players cash tournaments incredibly infrequently for each quartile.

We also notice that the x-axis is different for each plot, namely that the quartile with the highest maximum average time between cashes is 4900 days, whereas the other quartiles have similar maximum average times between cashes. From these plots, it is not clear which quartile of players cashes most frequently. We will determine this below.



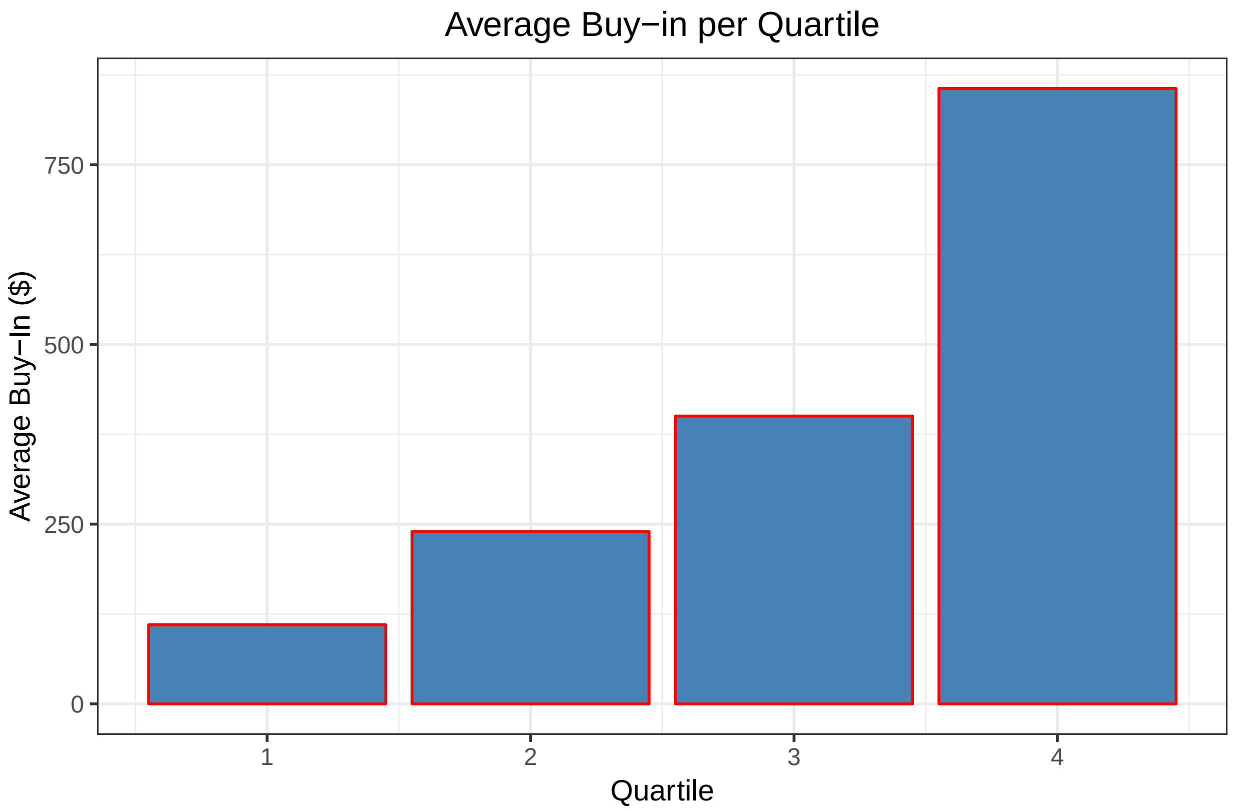
Source: The Hendon Mob

Naturally, we find that the players who cash the most frequently are those who have made the most money. Cashing frequently is certainly a good recipe for making money. However, it is important to note that this was not obvious.

It could have certainly been the case that players who have the least amount of cashes, cash just as frequently as those with high amounts of cashes. This would have indicated that players in quartile number 4 don't cash often, but when they do it's for large amounts of money.

Instead, we find large differences between quartiles 1 and 2 and quartiles 3 and 4. This plot seems to imply that players in quartiles 3 and 4 grind a similar amount as each other, but that maybe players in quartile 4 have achieved a lucky score here or there. To examine this hypothesis further, I turn to analyses of average buy-in amounts and lucky tournament scores (binks).

Average Buy-in per Quartile

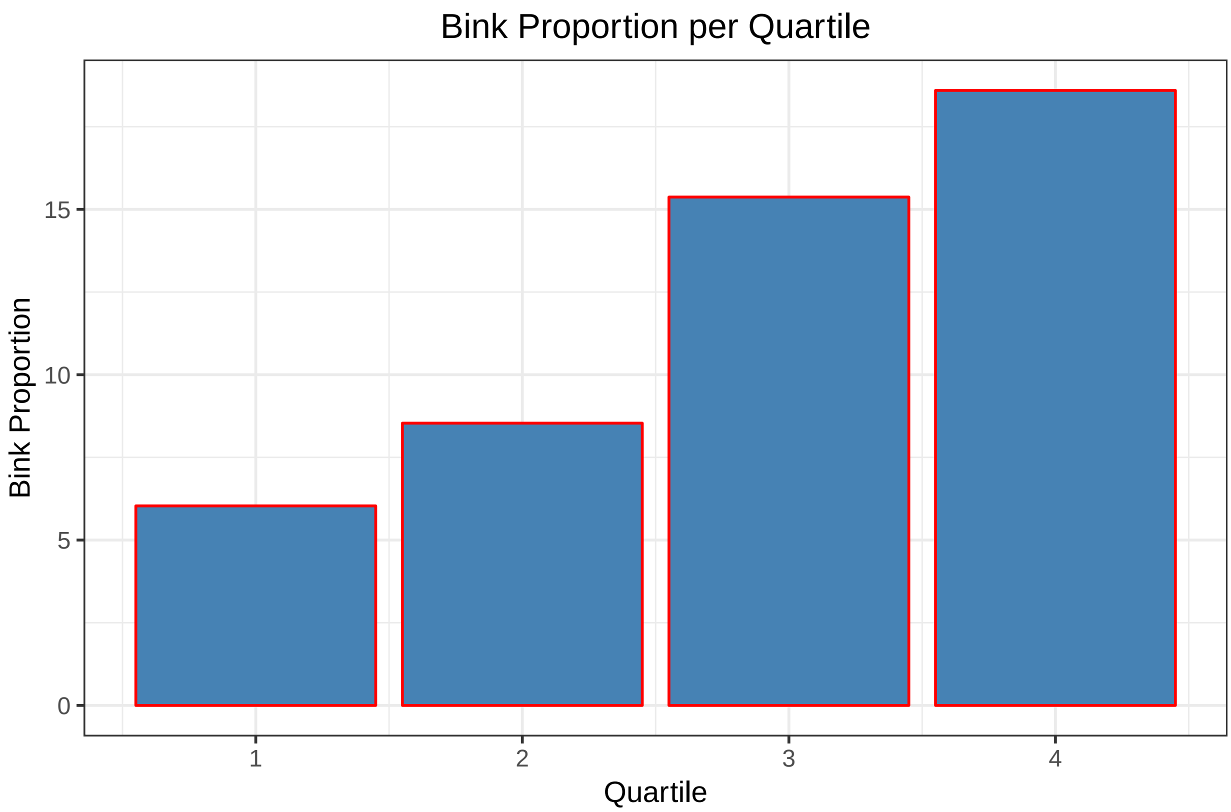


Source: The Hendon Mob

Players in quartile 4 play for higher stakes than the rest of the players in the database. Therefore, we can hypothesize that players in quartile 4 have cashed for more money not because they cash (or play) more frequently, but possibly because they play higher stakes than players in quartile 3. Is the key to having a high earnings number on the Hendon Mob simply to play high buy-in tournaments? Or is it also to get lucky in those tournaments? We will attempt to compare how often players get anamalous scores in each of the quartiles in the next section.

Binks per quartile

Here we take a first look at differences in binks per quartile, or basically trying to measure if players with different earning levels have experienced different levels of lucky scores.



Source: The Hendon Mob

We find that players with more cashes have a higher proportion of binks, meaning that they have achieved scores over 20x their average buy-in more frequently. It is quite possible that players who play higher stakes are more likely to have more opportunities for big scores, so there might be a confounding effect of average buy-in on bink proportion. We will attempt to understand which factors are actually responsible for earnings by running a regression model.

To summarize so far we've found the following through exploratory analysis:

- Positive influence on earnings
  - Average buy-in
  - Bink Proportion
- No influence on earnings
  - Distribution of time between cashes

Regression on Total Earnings/Summary

```
earnings_fit <- lm(sum_of_cashes ~ average_buy_in + average_time_btwn_cash + binks_proportion,
                   data = hendon_summaries_df)

summary(earnings_fit)
```

```
##
## Call:
```

```
## lm(formula = sum_of_cashes ~ average_buy_in + average_time_btwn_cash +
##     binks_proportion, data = hendon_summaries_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -554017  -21212  -12704   -7113 4201005
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8131.950   2801.092   2.903  0.00371 **
## average_buy_in           50.172      2.757  18.199  < 2e-16 ***
## average_time_btwn_cash  -15.541      7.248  -2.144  0.03206 *
## binks_proportion        204.505     77.199   2.649  0.00809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156300 on 5469 degrees of freedom
## Multiple R-squared:  0.05738,    Adjusted R-squared:  0.05687
## F-statistic:   111 on 3 and 5469 DF,  p-value: < 2.2e-16
```

The regression model shows that all three variables have a significant effect on earnings:

- Average buy-in: For every $1 **increase** in average buy-in, there is a corresponding $50 **increase** in earnings
- Average time between cashes: For every 1 day **increase** in average time between cashes, there is a a corresponding $15 **decrease** in earnings
- Binks Proportion: For every 1% **increase** in binks proprtion, there is a corresponding $204 **increase** in earnings

To summarize, the more people buy-in for, the more frequently they cash, and the more frequently they get lucky, the more money they make. Not the most surprising, but it's nice to see that the data reflect what would be intuitively true about poker earnings.

The size of the effects of the regression are also interesting because we find that anamalous tournament scores have the largest $ effect on earnings. It is difficult to compare the units of the predictor variables since a 1% increase in bink proportion is not equivalent to a 1 day increase in average time between cashes.

Nonetheless, it makes sense that having big scores relative to your buy-in level are going to be the most important influence on earnings. Binks not only directly influence earnings, but they also increase *earning potential.* Binks increase players' bankrolls and allow them to increase buy-ins so that they can make larger returns. Lesson being, luck is incredibly important, especially in tournament poker.