**Fantasy Premier League. Мастерство или удача?**

Меня этот вопрос волнует давно. Чтобы разобраться, я предпринял серьёзное исследование, по ходу которого пришлось преодолеть ряд технических трудностей. Краткий ответ звучит так: хороший результат в одном сезоне может быть получен благодаря удаче. Закономерно из сезона в сезон занимать высокие места могут лишь мастера. Давайте начнем с представления четырех выдающихся игроков Fantasy Premier League:

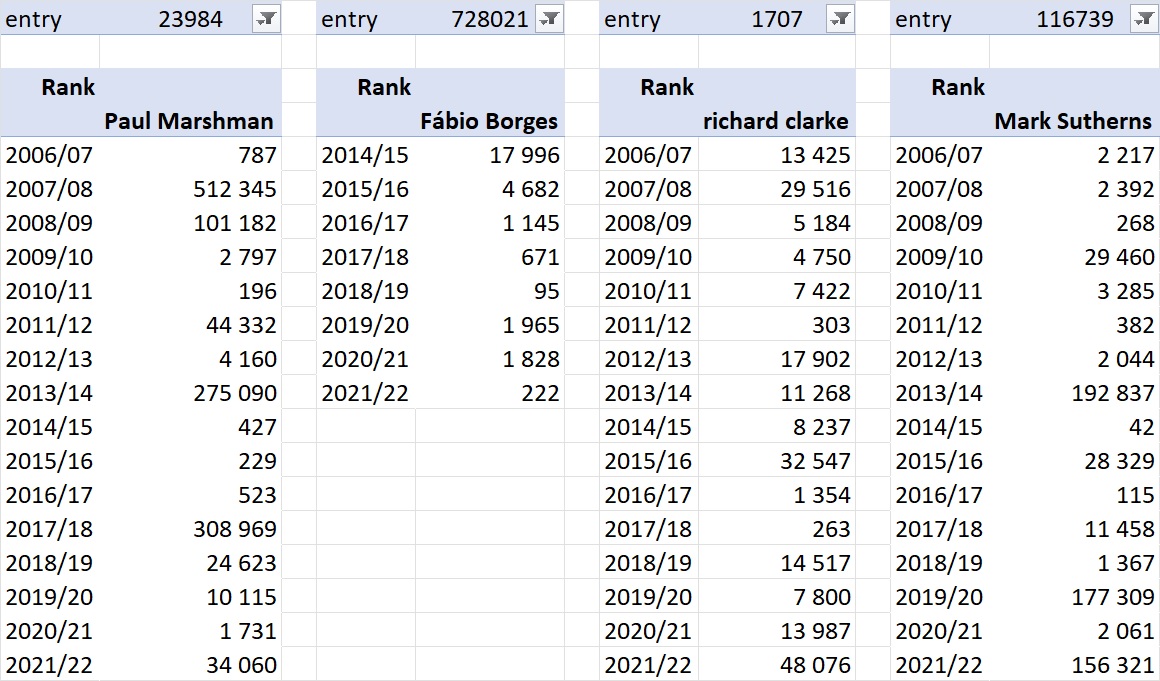


Рис. 1. Четыре потрясающих достижения: Paul Marshman наряду с тремя другими игроками 5 раз финишировал в ТОР-1k, Fábio Borges с большим отрывом от преследователей имеет самое низкое среднее значения ранга за последние пять сезонов, Richard Clarke единственный, кто во всех 16 сезонах финишировал в ТОР-50k, Mark Sutherns больше всех – 10 раз – финишировал в ТОР-5k

#### Завязка

Я увлекаюсь Fantasy Premier League, статистикой и Excel. Объединение этих интересов и позволило получить результаты, представленные в заметке.

Уже много лет я строю модели на основе больших данных FPL (см. [Лучшая команда Fantasy Premier League сезона 2017/18](http://baguzin.ru/wp/?p=18673)). Программирование не является моим коньком, и для скрапинга данных с сайта FPL я заказал программу у профессионалов. Но со временем условия изменились. Сначала [сайт](https://fantasy.premierleague.com) стал запрашивать подтверждение для входа, а затем потребовалось работать под VPN. В этих условиях скрапинг был затруднен.

Я не унывал, и начал гуглить. И удача! Нахожу [описание](https://buildmedia.readthedocs.org/media/pdf/fpl/latest/fpl.pdf) и [библиотеку](https://github.com/amosbastian/fpl) pyton для скрапинга сайта FPL. Прошу моего друга Вадима Миняйло помочь с pyton. После первых пробных попыток Вадим сообщает, что оказывается FPL предоставляет доступ к данным по API. Гуглю более прицельно, и нахожу [перечень](https://medium.com/@frenzelts/fantasy-premier-league-api-endpoints-a-detailed-guide-acbd5598eb19) точек доступа. Настраиваю запросы с помощью Power Query в Excel, и получаю массив данных – около 100М строк. Загружаю данные в модель Power Pivot и… комп не справляется. Покупаю новый в максимально доступной конфигурации – с процессором Intel Core i9-12900K, 3.20 GHz и оперативной памятью 32GB.

#### Методика

В рамках подготовки к сезону 2022/23 я прочитал несколько книг, и у [Пола Роджерса](https://baguzin.ru/wp/?p=25271) нашел ссылку на настоящее научное исследование опубликованное в рецензируемом электронном журнале [Джозеф Д. О'Брайен, Джеймс П. Глисон, Дэвид Дж.П. О'Салливан. Идентификация мастерства в Fantasy Premier League](https://baguzin.ru/wp/?p=25434). Исследование весьма интересное, но у меня к нему есть одно замечание. Авторы изучали ТОР-1М игроков по итогам сезона 2018/19. Проблема в том, что часть игроков из этой выборки попали на верхние строчки благодаря везению, а не подтвержденному мастерству. На мой взгляд интереснее выделить элитных игроков на основании многолетних результатов, и именно эту группу сравнивать с «поляной».

В цикле заметок я планирую представить вам интересные выводы о том, что отличает элитных игроков, а начну с выделения элиты.

Я скачал и проанализировал данные о истории выступлений 9,6М менеджеров, зарегистрировавшихся на сезон 2022/23 к 1 сентября 2022 г.

#### Статистика участий

Точные данных о числе участников за прошлые годы я не нашел, но в качестве оценки выбрал наихудший рейтинг в соответствующем сезоне среди менеджеров, зарегистрировавшихся на сезон 2022/23.

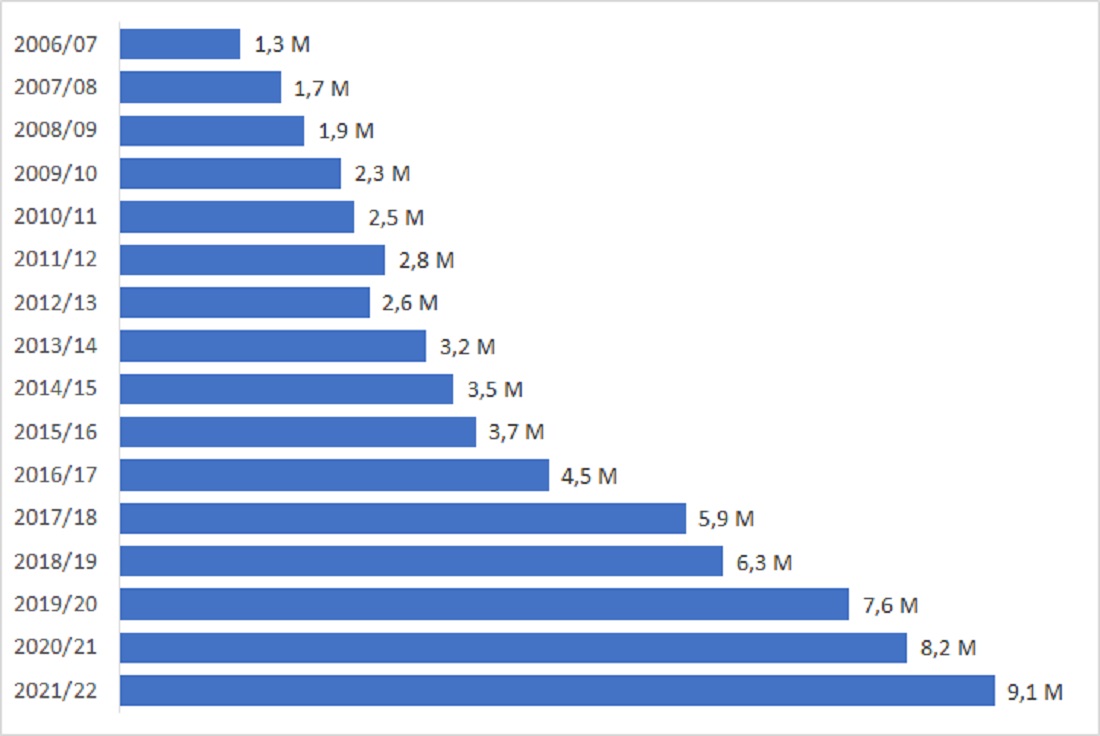


Рис. 2. За 16 сезонов количество аккаунтов увеличилось в 7 раз

Современную базу данных FPL ведет с сезона 2006/07. За более ранние периоды доступны лишь разрозненные данные.

Более трети аккаунтов – 3 464 000 – впервые зарегистрировались на сезон 2022/23. Но также есть и 52 тысячи менеджеров, которые приняли участие во всех 16 сезонах.

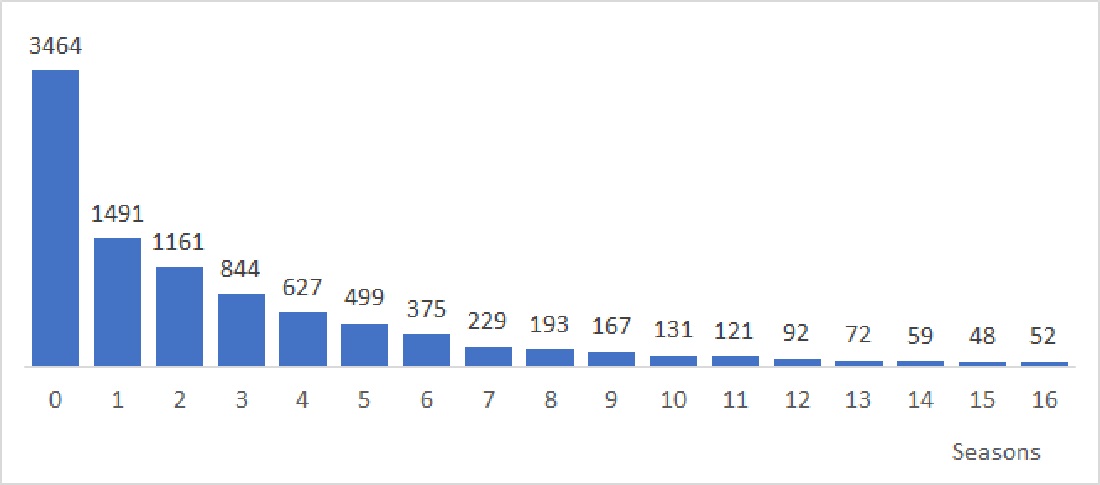


Рис. 3. Распределение аккаунтов по количеству участий в предыдущих сезонах, х1000

Естественно, что опытные игроки регистрируются раньше))

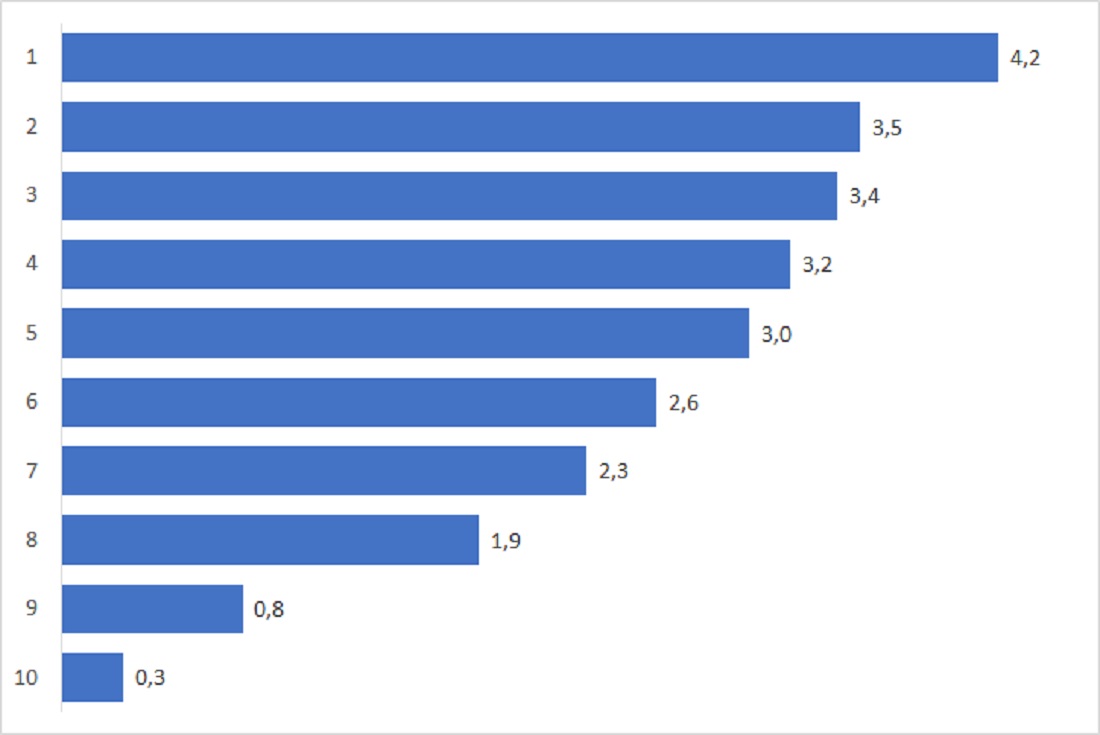


Рис. 4. Среднее число сезонов по группам аккаунтов; 1 по оси ординат соответствует аккаунтам с номерами от 1 до 1 000 000, 2 – аккаунты с номерами от 1 000 001 до 2 000 000 и т.д.

#### Элита

Чтобы определить, кого отнести к элите, мне предстояло ответить на ряд вопросов. Учитывать ли все выступления или только несколько последних? Отбирать элиту по абсолютным или относительным рейтингам? Ориентироваться на средние показатели или на число вхождений в ТОП? И наконец, сколько игроков отбирать? 100, 1000, 10 000?

Надо сказать, что элитные игроки представлены на ряде сайтов. Например, [livefpl.net](https://www.livefpl.net/), [fplstatistics](http://fplstatistics.co.uk/Home/Top50), [fplreview](https://fplreview.com/elite-data/), [myfplanalysis.co.in](https://myfplanalysis.co.in/) и [premierfantasytools.com](https://www.premierfantasytools.com/best-fpl-managers-list/). Я не думаю, что тот или иной список лучше или хуже. Важен общий принцип – высокие результаты на протяжении нескольких сезонов.

Чтобы не было перекоса в сторону игроков с больш***и***м числом участий, я решил считать результаты только за пять последних сезонов. Дабы отсечь случайные успехи, я не учитывал игроков с небольшим числом участий. За последние пять сезонов число аккаунтов увеличилось в полтора раза с 5,9М до 9,1М. С одной стороны, относительный рейтинг более справедлив. Ведь попасть в ТОП-10k при 5,9М участников легче, чем при 9,1М. С другой стороны, абсолютные рейтинги видны на сайте, и именно они фигурируют в обсуждениях. Что такое ТОП-10k понятно всем, а что такое ТОП-0,1%?.. На мой взгляд центральная мера (среднее) лучше отражает стабильность результатов, чем число попаданий в ТОП. И наконец медиана более робастна,[[1]](#footnote-1) чем среднее арифметическое, подверженное влиянию даже одного относительно неудачного сезона. Я включил в элиту 10 000 игроков:

1. 9318 лучших по медианному значению рейтинга, принявших участие во всех пяти сезонах 2017/2018 – 2021/22; медианное значение не более 55 000.
2. 537 лучших, принявших участие в четырех из пяти сезонов; медианное значение не более 40 000.
3. 145 лучших, принявших участие в трех из пяти сезонов; медианное значение не более 13 300.

Я решил на этом этапе отобрать в элиту 10k. А на последний важный вопрос, сколько игроков включать в элиту, ответить в рамках последующих исследований.

Чтобы оценить уровень игроков, посмотрите на показатели лучших и «худших» по категориям:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 5. Показатели игроков: 1) 1-й в категории «5 сезонов», 2) 1-й в категории «4 сезона», 3) 1-й в категории «3 сезона», 4) 9318-й в категории «5 сезонов», 5) 537-й в категории «4 сезона», 6) 145-й в категории «3 сезона»

Полный список элитных аккаунтов можно найти в приложенном Excel-файле. Также для дальнейшего исследования выделил три группы. Б***о***льшая группа включает меньшие.

Изображение выглядит как стол

Автоматически созданное описание

Рис. 6. Средняя медиана в сезонах 2017/2018 – 2021/22 и количество аккаунтов по элитным группам

#### Группа сравнения

Поскольку из тура в тур изучать все аккаунты (сейчас их более 10М) не представляется возможным, я выделил группу сравнения, и назвал ее *поляной*. Эту группу я сформировал на основе стратифицированной выборки:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 7. Правила формирования поляны

Такой выбор связан с том, что аккаунты с меньшими номерами формируются более опытными и вовлеченными менеджерами, поэтому стратифицированная выборка больше соответствует моей цели – сравнить игру элитного игрока и среднего. Среди аккаунтов с б***о***льшими номерами выше процент фейковых и случайных игроков.

Поскольку выбор поляны был случайным, в ней присутствует и некоторое число элитных игроков. Полный список аккаунтов, включенных в поляну, можно найти в приложенном Excel-файле.

В целом параметры выборок по всем участиям (а не только в сезонах 2017/2018 – 2021/22) выглядят следующим образом:

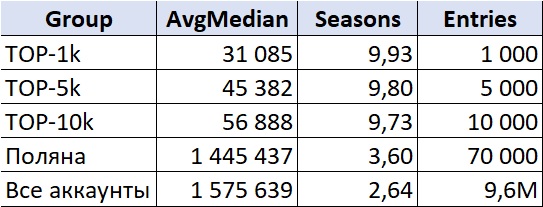


Рис. 8. Параметры выборок

1. [Роба́стность](https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C) — свойство статистического метода, характеризующее устойчивость к помехам. [↑](#footnote-ref-1)