

**Ilość asyst w meczach piłki nożnej, wyjaśniona za pomocą modelu ekonometrycznego**

**Projekt zaliczeniowy**

**Ekonometria  
Prowadząca: mgr Agnieszka Choczyńska  
Informatyka i Ekonometria  
rok 2, semestr letni**

**Autorzy:  
Bartosz Szczudło  
Rafał Mardyla  
Mateusz Ciepacz  
Oskar Skrzyniarz**

**Spis treści**

[1. Opis problemu oraz dobór danych 3](#_Toc105705118)

[2. Dopasowywanie zmiennych 5](#_Toc105705119)

[2.1. Uzupełnianie braków, dodawanie kolumn 5](#_Toc105705120)

[2.2. Wybór zmiennych kategorycznych 5](#_Toc105705121)

[2.3. Dodatkowe spostrzeżenia 7](#_Toc105705122)

[3. Podział danych 7](#_Toc105705123)

[4. Dobór zmiennych do modelu 9](#_Toc105705124)

# Opis problemu oraz dobór danych

Projekt ten ma za pomocą metod ekonometrycznych wyjaśnić ilość zdobywanych asyst

przez zawodników z pięciu głównych piłkarskich lig europejskich (tzw. *TOP 5*), tj. Premier League (Wielka Brytania), Serie A (Włochy), Bundesliga (Niemcy), La Liga (Hiszpania) oraz Ligue 1 (Francja). Wszelkie dane pochodzą ze strony *fbref.com*.

Do zbioru danych zaliczani są wszyscy piłkarze, którzy zostali zgłoszeni chociaż raz do kadry meczowej jakiegokolwiek zespołu z wymienionych wyżej lig w sezonie 2021/2022. Przy doborze statystyk, które mogą wpływać na ilość zdobytych bramek, sugerowaliśmy się własną wiedzą na temat piłki nożnej. W pierwszej kolejności wyselekcjonowane zostały statystyki dotyczące ilości oraz rodzajów podań wykonywanych przez zawodnika.[[1]](#footnote-1) Jest to dość oczywisty wybór, ponieważ asysta to również podanie. Następnie dodano zmienne zawierające ilość oddanych strzałów przez zawodnika.[[2]](#footnote-2) Ich wartość świadczy o tym czy dany piłkarz bierze często udział w akcjach bramkowych, a co za tym idzie czy przekłada się to na ilość zdobytych asyst. Dodatkowo zostały dodane statystyki dotyczące ilości wykreowanych akcji, prowadzących do oddania strzału lub zdobycia bramki[[3]](#footnote-3) oraz informacje personalne dotyczące samego zawodnika.

Wszystkie wyselekcjonowane dane zostały zaimportowane do pliku *database.csv[[4]](#footnote-4)*, który będzie wykorzystywany przy tworzeniu modelu ekonometrycznego. Plik posiada następujące kolumny:

* **#** - liczba porządkowa;
* **Player** – imię i nazwisko zawodnika.
* **Nation** – narodowość zawodnika.
* **Position** – pozycja na boisku.
* **Team** – zespół, w którym gra zawodnik
* **League** – liga, w której występuje piłkarz.
* **Age** – wiek piłkarza.
* **Born** – rok urodzenia zawodnika.
* **90s** – uśredniona ilość minut rozegranych przez zawodnika na jeden mecz.
* **TotPass** – ilość wszystkich podań wykonanych przez zawodnika (podanie musiało być celne, tj. piłka musiała dotrzeć do zawodnika tej samej drużyny).
* **PassAtt** – ilość wszystkich prób podań przez zawodnika. W zmienną wliczane są podania celne (tj. *Total Passes*) oraz podania niecelne.
* **TotDist** – sumaryczny dystans wszystkich podań celnych wykonanych przez zawodnika (podane w jardach).
* **PrgDist**– sumaryczny dystans wszystkich podań celnych wykonanych przez zawodnika w stronę bramki przeciwnika (podane w jardach).
* **ShortPass** – ilość krótkich podań (podanie na odległość poniżej 15 jardów) celnych wykonanych przez zawodnika.
* **ShortPassAtt** – ilość wszystkich prób krótkich podań przez zawodnika.
* **MedPass** – ilość średnich podań (podanie na odległość powyżej 15 jardów i poniżej 30 jardów) celnych wykonanych przez zawodnika.
* **MedPassAtt** – ilość wszystkich prób średnich podań przez zawodnika.
* **LongPass** – ilość długich podań (podanie na odległość powyżej 30 jardów) celnych wykonanych przez zawodnika.
* **LongPassAtt** – ilość wszystkich prób długich podań przez zawodnika.
* **Asissts** – liczba asyst wykonanych przez zawodnika (podanie do zawodnika, który bezpośrednio po tym podaniu zdobył bramkę).
* **Goals** – liczba bramek zdobytych przez zawodnika.
* **Shoots** – liczba wszystkich oddanych strzałów przez zawodnika (nie wliczane są rzuty karne).
* **SoT** – liczba wszystkich celnych strzałów oddanych przez zawodnika (nie wliczane są rzuty karne).
* **FreeKicks** – liczba strzałów oddanych z rzutów wolnych
* **SCA** – ilość sytuacji, które doprowadziły zawodnika lub kolegę z drużyny do strzału na bramkę rywala (tj. dryblingi, podania, strzały, faule).
* **PassLiveToShoot** – ilość podań wykonanych przez zawodnika z piłki toczącej się, po których bezpośrednio został oddany strzał.
* **PassDeadToShoot** – ilość podań wykonanych przez zawodnika z piłki stojącej, po których bezpośrednio został oddany strzał.
* **DribToShoot**– ilość skutecznie wykonanych dryblingów przez zawodnika, po których bezpośrednio został oddany strzał.
* **ShootToShoot** – ilość strzałów wykonanych przez zawodnika, po których bezpośrednio został oddany kolejny strzał.
* **FoulToShoot** – ilość przyznanych rzutów wolnych podyktowanych za przewinienie na zawodniku, po których bezpośrednio został oddany strzał.
* **DefToShoot** – ilość akcji defensywnych (np. przejęcie piłki), po których bezpośrednio został oddany strzał.
* **GCA** – ilość sytuacji, które doprowadziły zawodnika lub kolegę z drużyny do zdobycia bramki (tj. dryblingi, podania, strzały, faule).
* **PassLiveToGoal** – ilość podań wykonanych przez zawodnika z piłki toczącej się, po których bezpośrednio została zdobyta bramka.
* **PassDeadToGoal** – ilość podań wykonanych przez zawodnika z piłki stojącej, po których bezpośrednio została zdobyta bramka.
* **DribToGoal** – ilość skutecznie wykonanych dryblingów przez zawodnika, po których bezpośrednio została zdobyta bramka.
* **ShootToGoal** – ilość strzałów wykonanych przez zawodnika, po których bezpośrednio została zdobyta bramka.
* **FoulToGoal** – ilość przyznanych rzutów wolnych podyktowanych za przewinienie na zawodniku, po których bezpośrednio została zdobyta bramka.
* **DefToGoal** – ilość akcji defensywnych (np. przejęcie piłki), po których bezpośrednio została zdobyta bramka.

# Dopasowywanie zmiennych

W celu opracowania modelu ekonometrycznego niezbędne jest odpowiednie przygotowanie danych. Nie mogą one zawierać braków czy usuwanie zbędnych danych. Baza danych wykorzystywana w tym projekcie zawiera 2921 wierszy oraz 38 kolumn, z czego pierwsza kolumna jest liczbą porządkową, pięć kolejnych zawierają zmienne tekstowe, kolejne dwie zawierają informację na temat roku urodzenia i wieku, a pozostałe zawierają dane liczbowe, z czego jedna z nich (*90s*) jest ograniczona na przedziale [0, 90]. Reszta posiada jedynie dolną granicę równą 0.

## 2.1. Uzupełnianie braków, dodawanie kolumn

Na początku dokonujemy transformacji niektórych kolumn, aby ich wartości były przejrzyste i nie zawierały niepotrzebnych informacji.

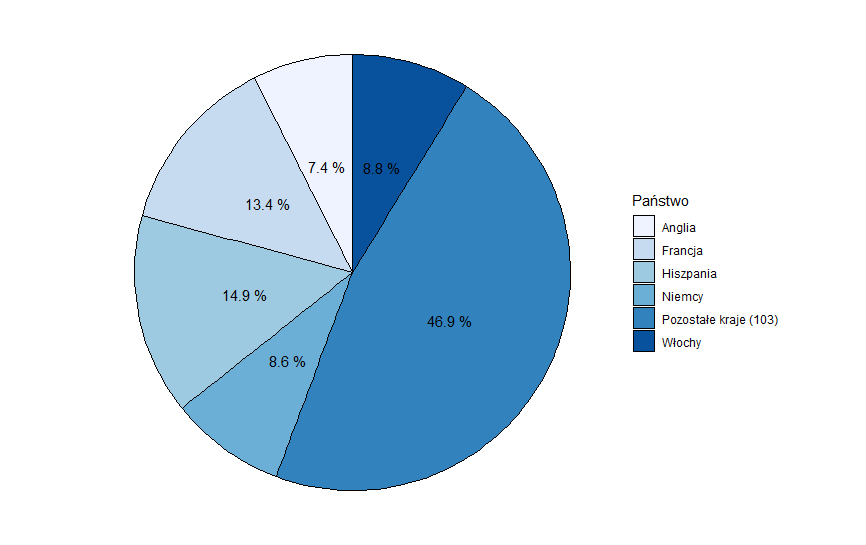
Poprzez komendę *summary* uzyskujemy dostęp do informacji na temat danych. Dzięki nim możemy stwierdzić, że jedyny występujący brak dotyczy roku urodzenia, wieku i kraju pochodzenia piłkarza w wierszu 1532. Jest to Javier Llabrés, urodzony w 2002 roku i pochodzący z Hiszpanii[[5]](#footnote-5). Ponieważ jest to pojedynczy brak, uzupełnimy go ręcznie.

Do danych dodana zostanie także nowa kolumna – **Formation**. Będzie ona przechowywała informację o przynależności danego piłkarza do konkretnej formacji. Bramkarzom oraz obrońcą zostanie przypisana formacja defensywna (*Def*), natomiast pomocnikom i napastnikom formacja ofensywna (*Off*).

## 2.2. Wybór zmiennych kategorycznych

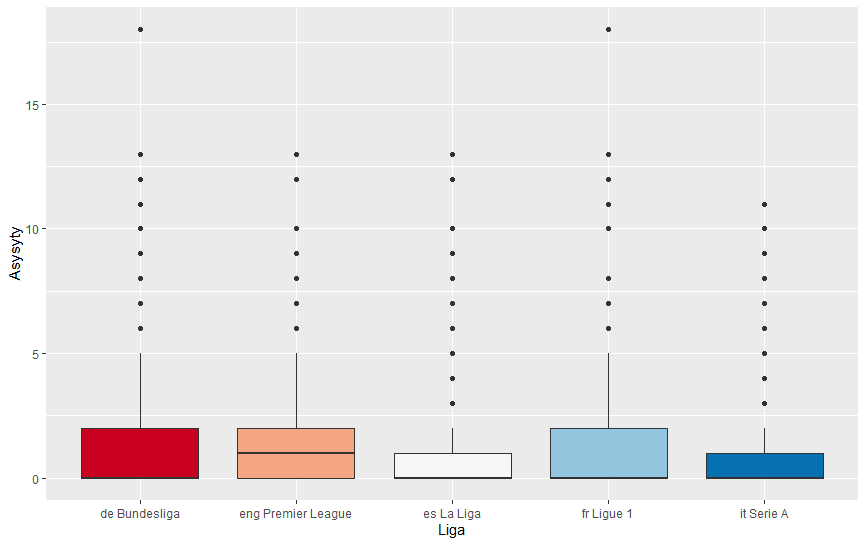
Dane zawierają teraz trzy zmienne kategoryczne. Są to kolejno: kraj pochodzenia (**Nation**), formację (**Formation**) oraz ligę (**League**). Taka ilość może niepotrzebnie skomplikować i wydłużyć proces modelowania, dlatego ograniczymy się tylko do jednej z nich. Na początku z uwagi na zbyt dużo kategorii i bardzo dużą rozbieżność, odrzucamy podział ze względu na kraj pochodzenia. Liczebność poszczególnych krajów jest zaburzona głównie przez przepisy juniorskie, które powodują, że obywatele Francji, Anglii, Hiszpanii, Niemiec czy Włoch są znacznie liczniejszą grupą od pozostałych krajów. Z poniższego wykresu można odczytać, że 53,1% wszystkich zawodników branych pod uwagę pochodzi z wyżej wymienionych pięciu krajów. Natomiast mniejszość stanowią piłkarze 103 pozostałych narodowości. Tak duże zaburzenie w proporcji pozwala nam na odrzucenie kraju pochodzenia jako wybranej zmiennej kategorycznej.

Liczebność reprezentantów poszczególnych krajów



Kolejno, możemy odrzucić kryterium ligi, w której dany piłkarz gra. Dane nie są zbytnio rozproszone (co można zaobserwować na poniższym wykresie pudełkowym). Od normy odstają przede wszystkim dwie ligi: hiszpańska i włoska. Dodatkowo w lidze angielskiej można zaobserwować wyższą wartość mediany. Powodem odrzucenia tej zmiennej jest jednak fakt, że w lidze niemieckiej jest mniejsza liczba klubów (18, w pozostałych po 20). Przez to zawodników z Bundesligi jest nieco mniej, a także mają oni o cztery spotkania mniej do rozegrania.

Rozbieżność asyst w poszczególnych ligach

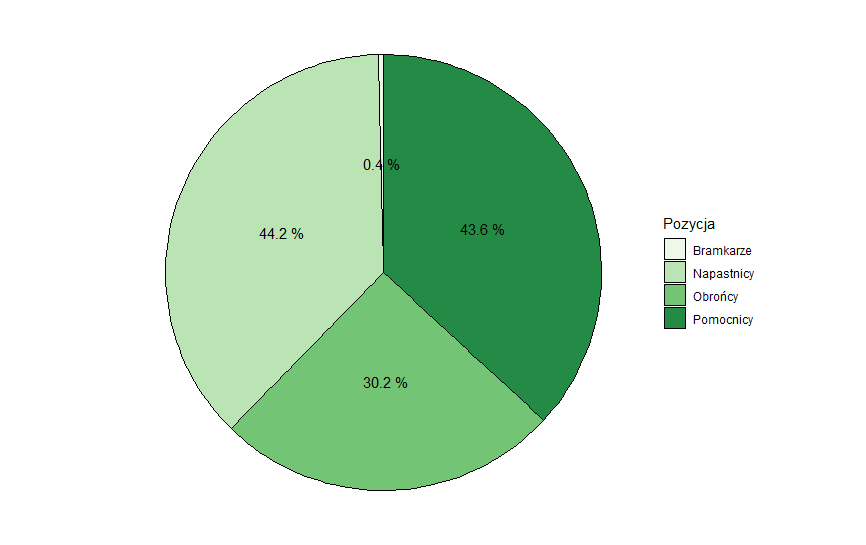


Jedynym pozostałym kryterium kategorycznym jest podział na formację na boisku. Jest to zmienna dychotomiczna, ponieważ przyjmuję wyłącznie dwie wartości. Ma ona także duże znaczenie, ponieważ spodziewamy się tego, że większość asyst zdobędą piłkarze grający w formacji ofensywnej, z uwagi na częstsze uczestnictwo w akcjach bramkowych. Z tych powodów, jako zmienną kategoryczną przyjmujemy podział na formację na boisku.

## 2.3. Dodatkowe spostrzeżenia

Analizując zależność ilości asyst pod względem pozycji piłkarza, możemy otrzymać ciekawy wniosek. Na sumaryczną ilość wszystkich asyst – 3460, tylko 13 zostało wykonanych przez bramkarzy, co daje nam 0,4% całości. Dlatego, aby nie zaburzyć zależności pomiędzy innymi zmiennymi, decydujemy się usunąć ze zbioru danych wszystkich zawodników na pozycji bramkarza. Takich piłkarzy jest 216.

Ilość asyst a pozycja na boisku



W celach uzyskania modelu należy również usunąć wszystkie zbędne kolumny. Dlatego do zmiennej **rawData** zapiszemy wcześniejszą bazę danych pomniejszoną o kolumny zawierające: imię i nazwisko piłkarza (**Player**), odrzucone zmienne kategoryczne (**Nation**, **League**), zespół zawodnika (**Team**), pozycję na boisku (**Possition**) oraz rok urodzenia (**Age**). Docelowa tabela z danymi zawiera więc 2705 obserwacji 32 zmiennych.

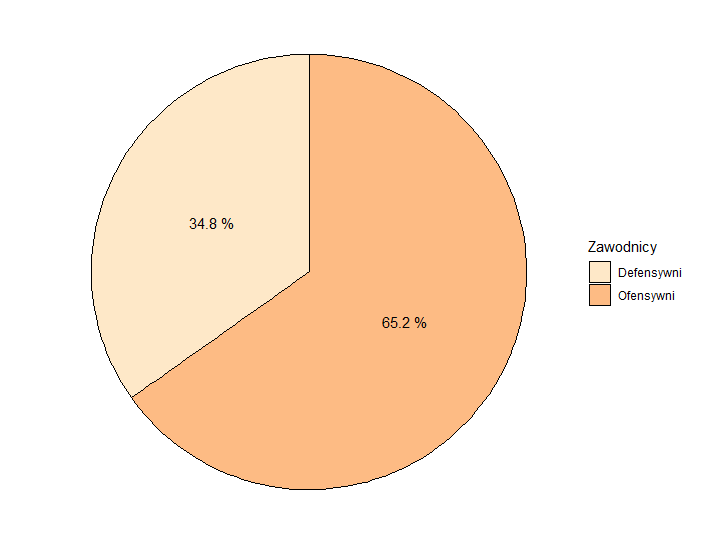
# Podział danych

Zebrane dane muszą zostać teraz podzielone na dwa zbiory: treningowy (uczący) oraz

testowy. Próba ucząca, to ta, na której zostanie wyprowadzony wzór, natomiast zbiór testowy służy do późniejszej weryfikacji całego modelu. Dlatego przyjmijmy założenie, że pierwszy z nich będzie zawierał 70% wszystkich obserwacji, drugi natomiast pozostałe 30%. Daje nam to podział 1893 zawodników w próbie uczącej oraz 812 w próbie testowej.

Przed podziałem zbadamy procentowy podział całego zbioru ze względu na formację zawodnika. Dostajemy zależność, że 65,2% (1763) zawodników należy do formacji ofensywnej. Do defensywnej natomiast – 34,8% (942). Żeby jak najlepiej oddać prawidłowość predykcji, zadbamy o podobną proporcję w obu podzbiorach (65% - 35%).

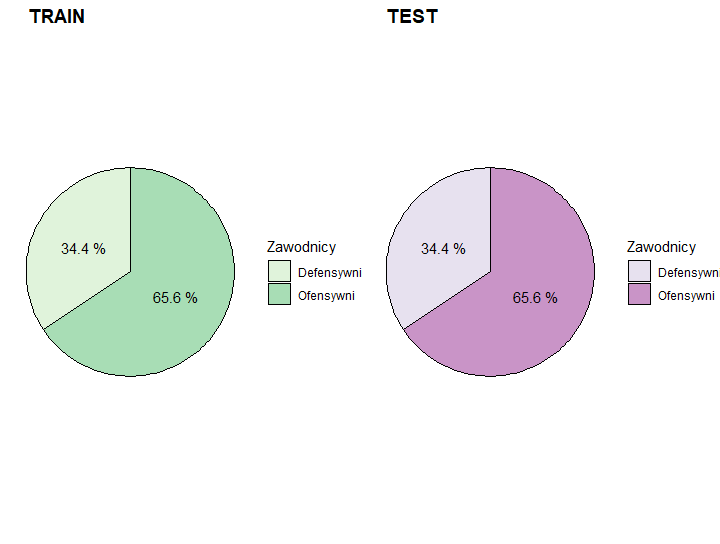
Podział zawodników na formację



Aby tego dokonać podzielimy *rawData* na dwa podzbiory zawierające piłkarzy tylko z jednej formacji. Następnie obliczymy ilu piłkarzy z każdej z nich powinno znaleźć się w zbiorze próby uczącej. Otrzymujemy wynik: 1230 piłkarzy ofensywnych i 663 defensywnych. Następnie wykorzystując bibliotekę *dplyr* wybierzemy z otrzymanych wcześniej podzbiorów odpowiednią ilość piłkarzy z każdej formacji. Wybór piłkarzy jest losowy.

Otrzymane w ten sposób dwa zbiory zostaną naszą próbą uczącą (**train**) i testową (**test**). Dzięki temu w obu zbiorach otrzymujemy zbliżone proporcję obu formacji wynoszące odpowiednio 65,6% (ofensywni) do 34,4% (defensywni). Na koniec możemy z obu podzbiorów usunąć kolumnę, zawierającą liczbę porządkową (**#**), którą wykorzystaliśmy przy podziale.

Proporcję podziału formacji



# Dobór zmiennych do modelu

## 4.1. Wybór metody selekcji

Przy selekcji zmiennych posłużymy się metodą eliminacji w oparciu o wartość skorygowanego współczynnika determinacji. Usuwamy tą zmienną, która posiada największe prawdopodobieństwo testowe (*p-value*). Jeżeli współczynnik R2 zwiększył się lub pozostał na tym samym poziomie – operację powtarzamy. Jeżeli zmniejszył się, powracamy do poprzedniej ilości zmiennych i obieramy otrzymaną wartość współczynnika za maksymalną, a pozostałe zmienne uznajemy za najodpowiedniejsze dla modelu.

## 4.2. Selekcja zmiennych

Model wyjściowy posiada 31 zmiennych oraz współczynnik R2 równy 84,62%. Jako dwie pierwsze zmienne odrzucamy zmienne przechowujące informację o liczbę akcji defensywnych prowadzących do gola bądź strzału (**DefToShoot**, **DefToGoal**) z uwagi na brak wystarczającej ilości informacji do użycia w modelu. Po usunięciu wartość współczynnika pozostaje na tym samym poziomie.

Następnie wybieramy zmienną o najwyższym *p-value*. W tym przypadku jest to ilość strzałów prowadzących do zdobycia gola (**ShootToGoal**). Po usunięciu warte ość R2 wzrosła do 84,63% - selekcję prowadzimy dalej.

Kolejną usuniętą zmienną z *p-value* wynoszącym 0.784362 jest dystans progresywnych podań (**PrgDist**). Wartość współczynnika determinacji ponownie wzrosła, tym razem do 84,64%. Operację powtarzamy, usuwając uśrednioną liczbę rozegranych minut w jednym meczu (**x90s**). R2 po usunięciu zostało na tym samym poziomie. Wartość współczynnika nie zmienia się również po usunięciu wartości ilości fauli prowadzących do zdobycia gola (**FoulToGoal**) o *p-value* równym 0.185183.

Dopiero przy usunięciu zmiennej odpowiedzialnej za ilość fauli prowadzących do strzału (**FoulToShoot**) spada wartość R2 do poziomu 84,63%. Dlatego ponownie dodajemy wyżej wymienioną zmienną.

## 4.3. Analiza selekcji

Po odrzuceniu niektórych zmiennych, maksymalna wartość skorygowanego współczynnika determinacji wynosi 84,64%. Model powinien posiadać zatem 25 zmiennych.

ZAPOMNIAŁEM O TYM ŻE ZBIORY TESTOWE I TRENINGOWE DZIELĄ SIĘ LOSOWO xD. DLATEGO PRZY KAŻDYM PONOWYM URUCHOMIENIU SELEKCJA BĘDZIE SIĘ RÓŻNIŁA. TRZEBA ZADECYDOWAĆ CZY URUCHAMIAMY RAZ PODZIAŁ I JUŻ WIĘCEJ GO NIE ZMIENAMY CZY DZIELIMY PRÓBĘ W NIELOSOWY SPOSÓB.

FRAGMENT DO DOKOŃCZENIA PO DECYZJI

1. Statystyki pochodzą ze strony: *https://fbref.com/en/comps/Big5/passing/players/Big-5-European-Leagues-Stats* (data dostępu: 03.06.2022). [↑](#footnote-ref-1)
2. Statystyki pochodzą ze strony: *https://fbref.com/en/comps/Big5/shooting/players/Big-5-European-Leagues-Stats* (data dostępu: 03.06.2022). [↑](#footnote-ref-2)
3. Statystyki pochodzą ze strony: *https://fbref.com/en/comps/Big5/gca/players/Big-5-European-Leagues-Stats* (data dostępu: 03.06.2022). [↑](#footnote-ref-3)
4. Wszystkie statystyki zostały odseparowane z tabel z witryny *fbref.com*, a następnie wklejone w jeden plik *csv*. [↑](#footnote-ref-4)
5. Informacja pochodzi ze strony: *https://www.transfermarkt.pl/javier-llabres/profil/spieler/706138* (data dostępu: 06.06.2022). [↑](#footnote-ref-5)