

# Raport 1

Martyna Bielec, Maciej Karczewski

2022-11-23

## Wprowadzenie

Strona internetowa Lichess jest drugą co do wielkości platformą na świecie, która umożliwia grę w szachy online. Użytkownicy mogą grać z komputerem lub między sobą. Każdy gracz ma przypisany ranking (początkowo ma on wartość 1500), który jest ustalany na podstawie jego umiejętności. Jeśli użytkownik rozgrywa partię “rankingową”, w jej wyniku jego ranking rośnie lub maleje.

Dane, które zostaną poddane analizie, zawierają informacje o ponad 20 000 rozgrywkach szachowych ze strony Lichess Games. Dostępne są na witrynie kaggle. Składają się z następujących kolumn:

- id (zawiera elementy typu “character”) - numer ID rozgrywki
- rated (“character”) - wartość “True” lub “False” w zależności od tego, czy partia jest “rankingowa”
- created\_at (“numeric”) - czas rozpoczęcia gry
- last\_move\_at (“numeric”) - czas zakończenia gry
- turns (“numeric”) - liczba rozegranych tur
- victory\_status (“character”) - status zakończenia gry: “mate” dla zakończenia matem, “resign” dla zakończenia przez poddanie się jednej ze stron, “draw” dla remisu lub “outoftime” dla rozgrywki zakończonej z powodu upływu czasu
- winner (“character”) - zwycięzca gry: “white”, jeśli wygrał gracz biały, “black”, jeśli czarny lub “draw” w przypadku remisu
- increment\_code (“character”) - składa się z dwóch liczb rozdzielonych znakiem “+”; pierwsza oznacza liczbę minut przypadających na gracza na wszystkie jego ruchy, druga liczbę sekund, o którą zwiększa się czas gracza po wykonaniu ruchu
- white\_id (“character”) - ID gracza białego
- white\_rating (“numeric”) - ranking gracza białego
- black\_id (“character”) - ID gracza czarnego
- black\_rating (“numeric”) - ranking gracza czarnego
- moves (“character”) - wszystkie ruchy zapisane w standardowej notacji szachowej
- opening\_eco (“character”) - standaryzowany kod dla rozgranego otwarcia
- opening\_name (“character”) - nazwa rozgranego otwarcia
- opening\_ply (“character”) - liczba ruchów w fazie otwarcia

Celem analizy będzie odpowiedzenie na pytanie - jakie jest prawdopodobieństwo różnych zakończeń gry: wygranej białego gracza, czarnego lub remisu? Będziemy rozważać ten problem pod kątem rozgranego otwarcia, rankingu graczy, różnicy w rankingu między graczem białym i czarnym oraz liczby tur w partii. Użyjemy metody bootstrap do stworzenia przedziałów ufności estymowanego prawdopodobieństwa.

## Obróbka danych

### Usunięcie niepotrzebnych kolumn

Nasze dane zawierają dużo kolumn z, których nie korzystamy dlatego usuniemy niepotrzebne kolumny w wyniku czego nasza tabela z danymi wygląda następująco.

	turns	winner	white_rating	black_rating	opening_name
## 1	13	white	1500	1191	Slav Defense: Exchange Variation
## 2	16	black	1322	1261	Nimzowitsch Defense: Kennedy Variation
## 3	61	white	1496	1500	King's Pawn Game: Leonaridis Variation
## 4	61	white	1439	1454	Queen's Pawn Game: Zukertort Variation
## 5	95	white	1523	1469	Philidor Defense
## 6	5	draw	1250	1002	Sicilian Defense: Mongoose Variation

### Usunięcie duplikatów i wierszy z brakiem danych

W następnym kroku usuwamy duplikujące się wiersze oraz wiersze w których występuje brak danych

### Zastosowanie One Hot Encoding dla kolumny “winner”

Wynik naszej partii jest zawarty w kolumnie “winner”. Ta kolumna ma 3 wartości: “white”, “black” i “draw”. Dla ułatwienia analizy stworzymy 3 dodatkowe kolumny o odpowiednich nazwach “white\_win”, “black\_win”, “draw”. Każda z tych kolumn ma wartość 1 lub 0. Gdy w kolumnie “winner” jest “white” to 1 przypisujemy do “white\_win”, jeśli “black” to do “black\_win” natomiast jeśli “draw” to do “draw”. W przeciwnych przypadkach przypisujemy 0.

```
df <- df %>% mutate(  
  white_win = ifelse(winner == 'white', 1, 0),  
  black_win = ifelse(winner == 'black', 1, 0),  
  draw = ifelse(winner == 'draw', 1, 0)  
)
```

### Otwarcia (?)

W raporcie przeanalizujemy prawdopodobieństwo różnych zakończeń rozgrywki także ze względu na wybrane otwarcie. Aby uzyskać rzetelne wyniki, nie możemy uwzględnić otwarć, które rozegrała zbyt mała liczba graczy. Okazuje się jednak, że kolumna opening\_name zawiera 1477 unikalnych wartości. Dzieje się tak, ponieważ dane rozróżniają wiele wariantów poszczególnych otwarć. Przykładowo “obrona sycylijska” została rozegrana w 181 różnych wariantach, takich jak:

```
## [1] "Sicilian Defense: Mongoose Variation"  
## [2] "Sicilian Defense: Bowdler Attack"  
## [3] "Sicilian Defense: Smith-Morra Gambit #2"  
## [4] "Sicilian Defense: Canal Attack | Main Line"  
## [5] "Sicilian Defense: Dragon Variation | Yugoslav Attack | Main Line"
```

W celu umożliwienia analizy otwarć stworzyliśmy nową kolumnę, zawierającą jedynie nazwę otwarcia (bez uwzględnienia wariantu). Nazwaliśmy ją “openings\_general”. Zawiera ona 149 wartości unikalnych. Przeanalizujemy te otwarcia, których zostało rozegranych więcej niż 500, czyli:

```
## # A tibble: 10 x 2
##   openings_general    total_count
##   <chr>              <int>
## 1 Sicilian Defense    2581
## 2 French Defense      1377
## 3 Queen's Pawn Game   1203
## 4 Italian Game         952
## 5 King's Pawn Game    897
## 6 Ruy Lopez           832
## 7 English Opening      715
## 8 Scandinavian Defense 707
## 9 Philidor Defense     670
## 10 Caro-Kann Defense   584
```

gdzie kolumna “total\_count” oznacza liczbę rozegranych gier z wykorzystaniem danego otwarcia.

## Analiza danych

### Prawdopodobieństwo różnych zakończeń rozgrywki bez warunkowania.

W celu rozważenia od czego zależy prawdopodobieństwo różnych zakończeń rozgrywki, najpierw oszacowaliśmy to prawdopodobieństwo w ogólności. Liczba rozegranych partii wynosi:

```
## [1] 19618
```

Z czego gracz biały wygrał następującą liczbę razy:

```
## [1] 9782
```

Natomiast czarny:

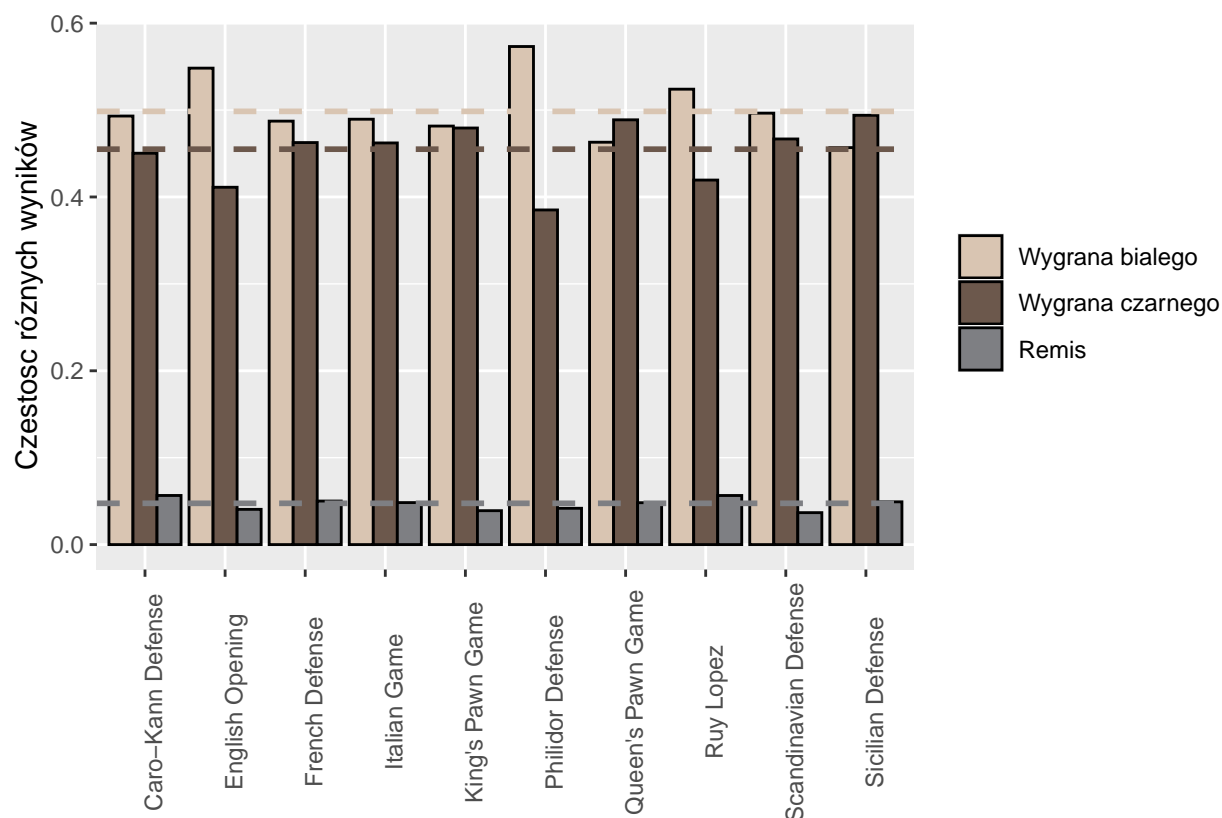
```
## [1] 8916
```

Wynika z tego, że szacowane prawdopodobieństwo wygrania gracza białego wynosi w przybliżeniu 49,85%, gracza czarnego 45,40%, a remisu 4,75%.

W dalszej części analizy sprawdzimy, jak to prawdopodobieństwo się zmienia, jeśli odpowiednio uwarunkujemy rozgrywkę.

### Warunkowanie otwarciem

Przeanalizujemy jak zmienia się prawdopodobieństwo różnych wyników partii w zależności od otwarcia. Pod uwagę wzięliśmy dziesięć najpopularniejszych, a dla każdego z nich dysponowaliśmy ponad 500 obserwacjami. Obliczyliśmy częstość wygranych oraz remisu dla poszczególnych otwarć:

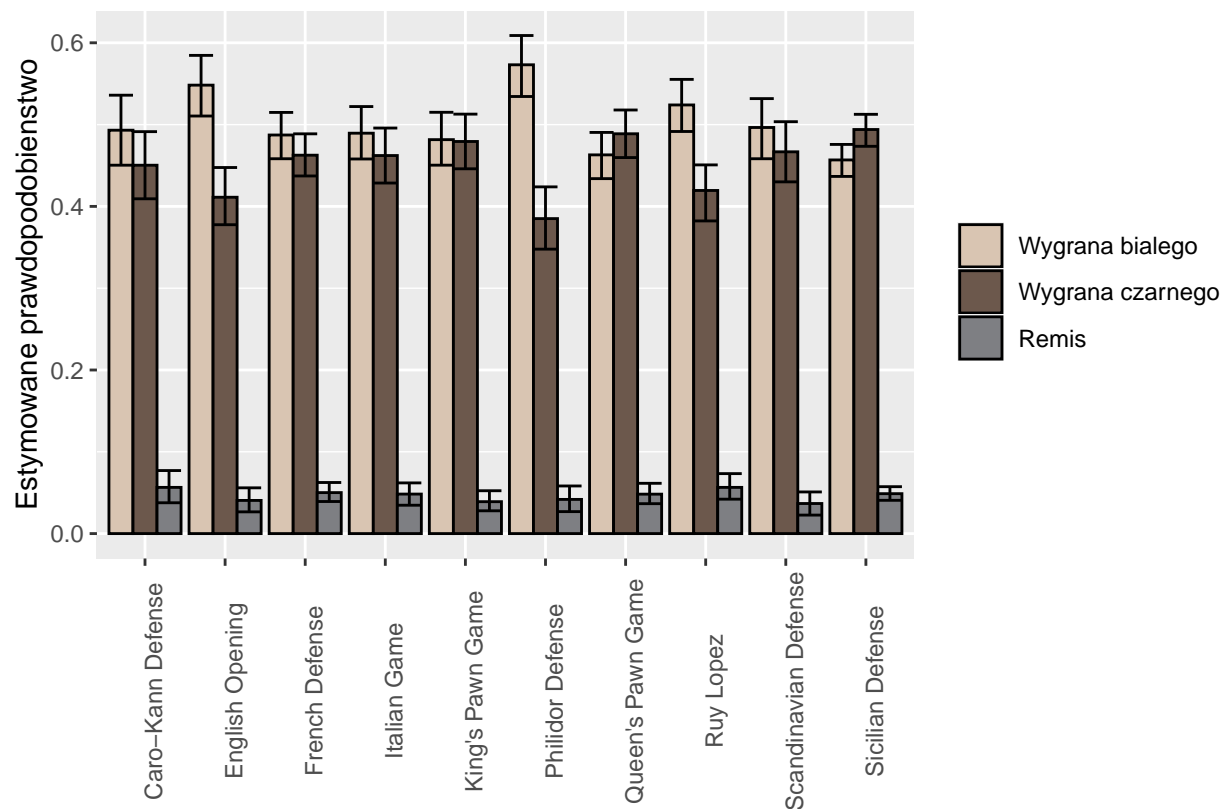


Przerzywanymi liniami oznaczyliśmy częstość poszczególnych wyników bez warunkowania: beżowa linia to częstość wygrywania gracza białego, brązowa czarnego, a szara remisu.

Z powyższego wykresu wynika, że biały gracz statystycznie częściej niż dla średniej rozgrywki wygrywa dla obrony Philidora ("Philidor Defense"), Otwarcia Angielskiego ("English Opening") oraz partii hiszpańskiej ("Ruy Lopez"). Szanse czarnego gracza na wygraną zwiększają się natomiast przy rozegraniu pozostałych z analizowanych otwarć oprócz obrony Caro-Kann ("Caro-Kann Defense"), gdzie zmalała jednocześnie częstość wygrywania gracza białego, natomiast zwiększyła się częstość remisu.

Przy dwóch z analizowanych otwarć gracz czarny nie wygrywał częściej niż gracz biały: przy obronie sycylijskiej ("Sicilian Defense") oraz gry pionkiem hetmańskim ("Queen's Pawn Game").

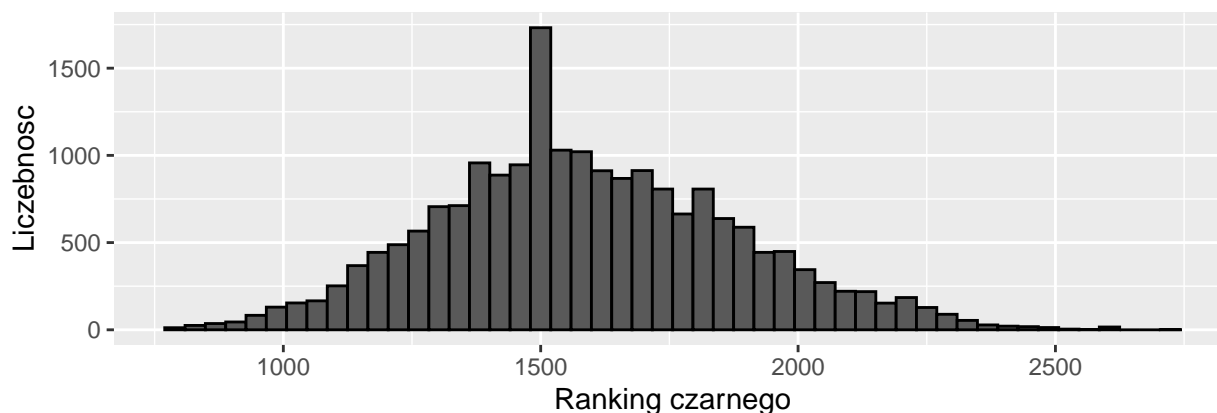
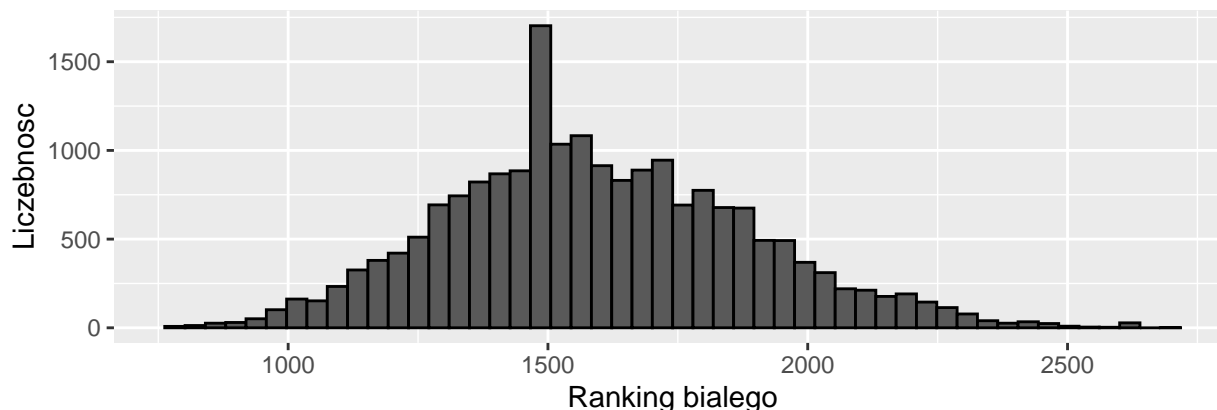
Aby sprawdzić dokładność naszych obliczeń oraz wysunąć wnioski z danych, stworzyliśmy przedziały ufności za pomocą metody bootstrap na poziomie ufności równym 0,05.



Okazuje się, że na poziomie ufności 0,05 nie możemy stwierdzić dla każdego otwarcia, prawdopodobieństwo wygrania którego gracza jest większe. Konkretnie wnioski możemy wysunąć jednak dla Otwarcia Angielskiego, Obrony Philidora oraz partii hiszpańskiej - dla nich stwierdzamy na poziomie ufności 0,05, iż prawdopodobieństwo wygrania białego gracza jest wyższe niż prawdopodobieństwo wygrania czarnego gracza. Prawdopodobieństwo remisu nie różni się natomiast znacząco dla żadnego otwarcia.

### Warunkowanie rankningiem białego

Gracze mają różne poziomy rankigu, więc sprawdzimy jak wyglądają prawdopodobieństwa różnych zakończeń rozgrywki w zależności od rankingu białego gracza. Na początku zobaczymy, jak wygląda rozkład rankingu białego gracza a jak czarnego.

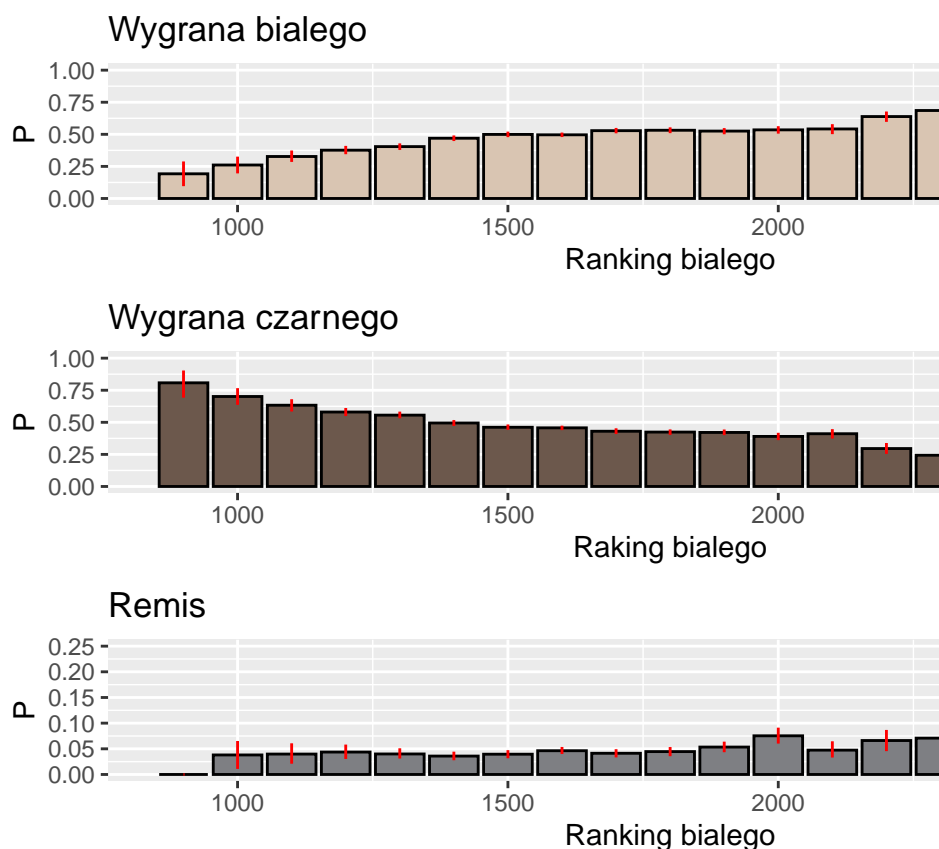


Wygenerowane histogramy sugerują, że rozkład rankingu białego i czarnego gracza są bardzo podobne. Rodzi się pytanie, jak wyglądają ich średnia, mediana, wariancja, skośność czy kurtoza.

	biały	czarny
średnia	1596,15	1588,39
mediana	1567	1562
wariancja	84221,18	84214,67
skośność	0,30	0,25
kurtoza	0,029	-0,052

Analizując statystyki opisowe możemy dojść do wniosku, że biały gracz ma średnio większy ranking. Różnica między średnimi wynosi tylko około 8 punktów rankingowych, co przy średnim rankingu nieco mniejszym niż 1600 jest znikomą przewagą. Dodatkowo możemy odczytać, że mediana rankingu białego gracza jest o 5 punktów większa od mediany czarnego, co znowu jest małą różnicą przy medianie powyżej 1560. Możemy zobaczyć także biały ma znowu minimalnie większą wariancję ale różni się tylko o 7 co przy wariancji białego równej 84221.18 nie jest żadną różnicą. Następnie możemy załważyć, że rozkład rankingu czarnego ma mniejszą skośność ale obydwa rozkłady mają na tyle małe skośności w stosunku do odchylenia standardowego, że możemy przyjąć że ich rozkłady są symetryczne. Jeśli chodzi o kurtozę to wyniki jej są na tyle blisko 0, że możemy przyjąć, że rozkłady mają tyle samo danych odstających co rozkład normalny.

Z racji, że rankig jest zmienną ciągłą to podzielimy rankig białego na przedziały np [1300;1400) i taki przedział będziemy ozacznać jako 1300. Każdy przedział ma szerokość 100. Dodatkowo usuwamy przedziały gdzie liczba gier wynosi mniej niż 11 Sprawdzimy teraz jak wyglądają Prawdopodobieństwa końcowych wyników w zależności od przedziału rankingu białego. Na wykresie czerwone krski pionowe to są przedziały

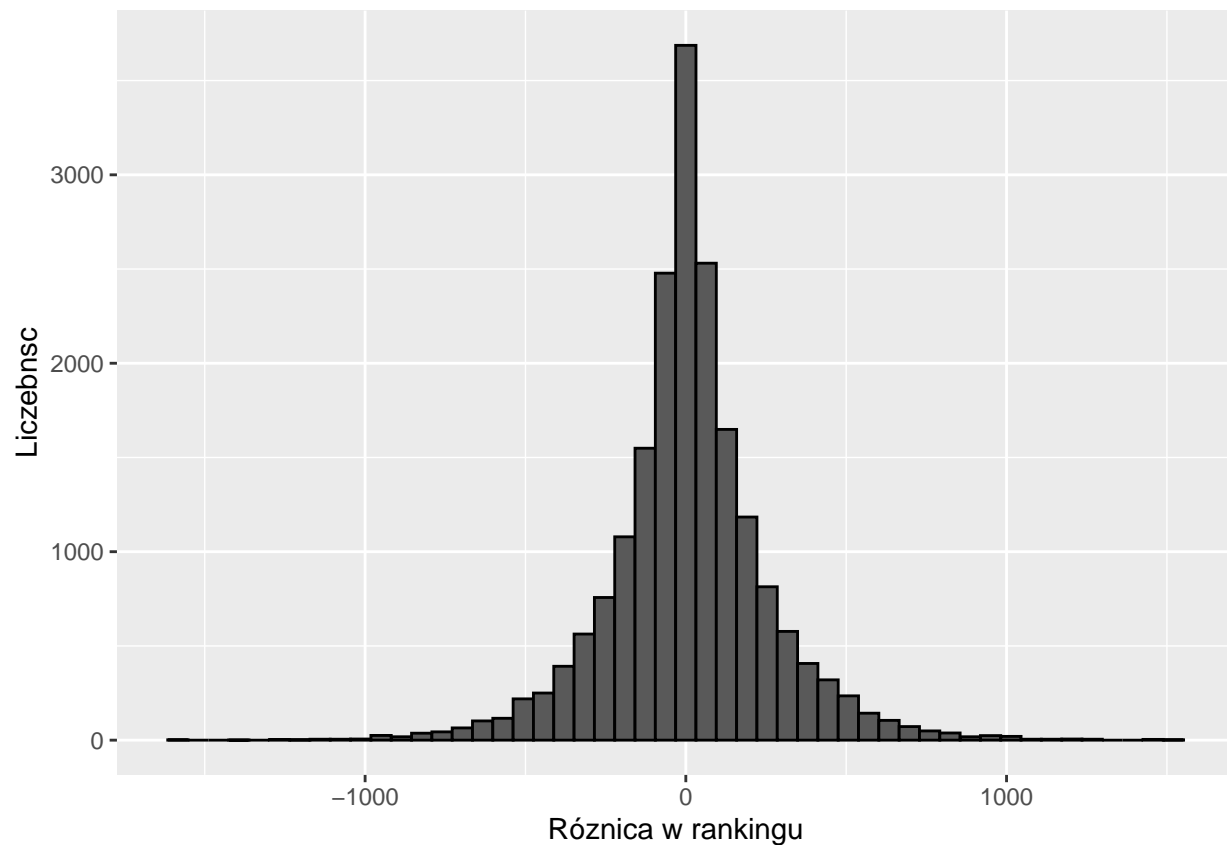


ufości dla prawdopodobieństwa wygranej.

Analizując wykres możemy załważyć, że waz ze wzrostem rankingu białego rośnie jego prawdopodobieństwo wygranej, natomiast maleje prawdopodobieństwo wygrania czarnego. Możemy też odczytać że prawdopodobieństwo remisu rośnie, ale może to być niekoniecznie stwierdzenie prawdziwe ponieważ patrząc na średnie można tak wnioskować natomiast patrząc na przedziały ufości średniej niekoniecznie ponieważ przy wysokich rankingach się zawierają w siebie więc lepiej stwierdzić, że nie wiadomo jaki ma wpływ.

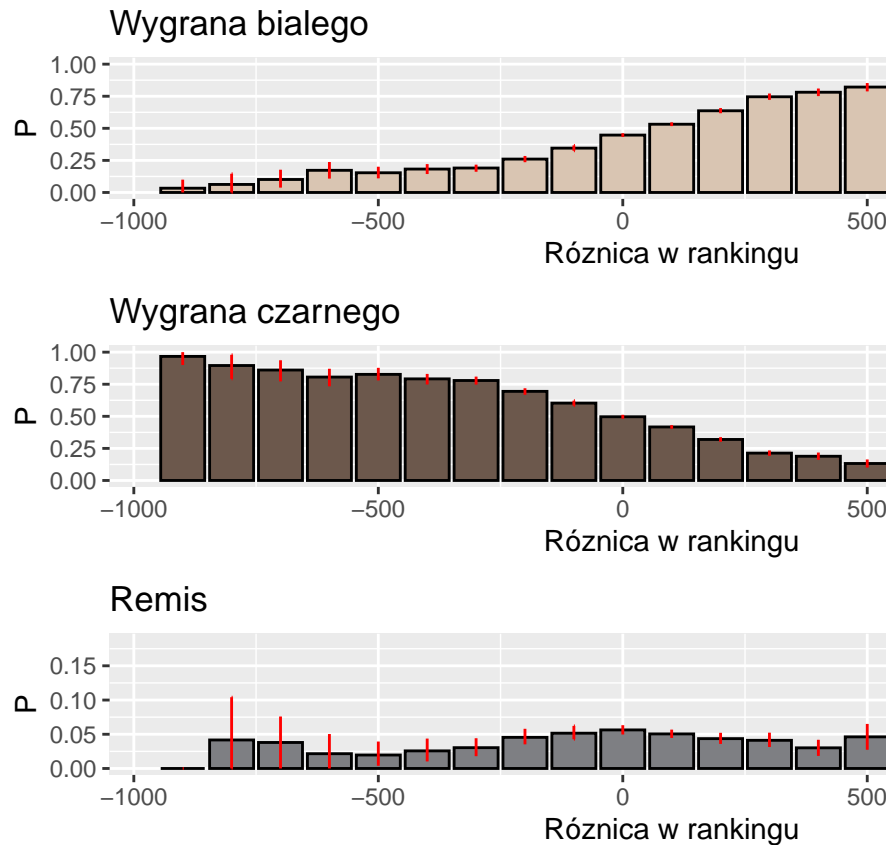
### Warunkowanie różnicą rankingu białego i czarnego gracza

Analizowaliśmy jak wyglądają prawdopodobieństwa zakończenia parti od rankingu białego, ale ranking czarnego powinien mieć też wpływ więc Dodamy nową kolumnę, która będzie oznaczać różnice między rankingiem białego i czarnego. Rozkład różnic w rankingu wygląda następująco:



Jak możemy zobaczyć rozkład różnic w rankingu jest niemal symetryczny a średnią wartość jest w okolicach 0 a dokładnie wynosi 7.56. Możemy zobaczyć, że wartość różnic rankingu zawierają się w przedziale  $[-1000, 1000]$  więc podzielimy różnice na kategorię o długości 100 tak jak w przypadku rankingu białego. Zobaczmy teraz jak wyglądają prawdopodobieństwa zakończenia partii w zależności od przeziału różnic dla



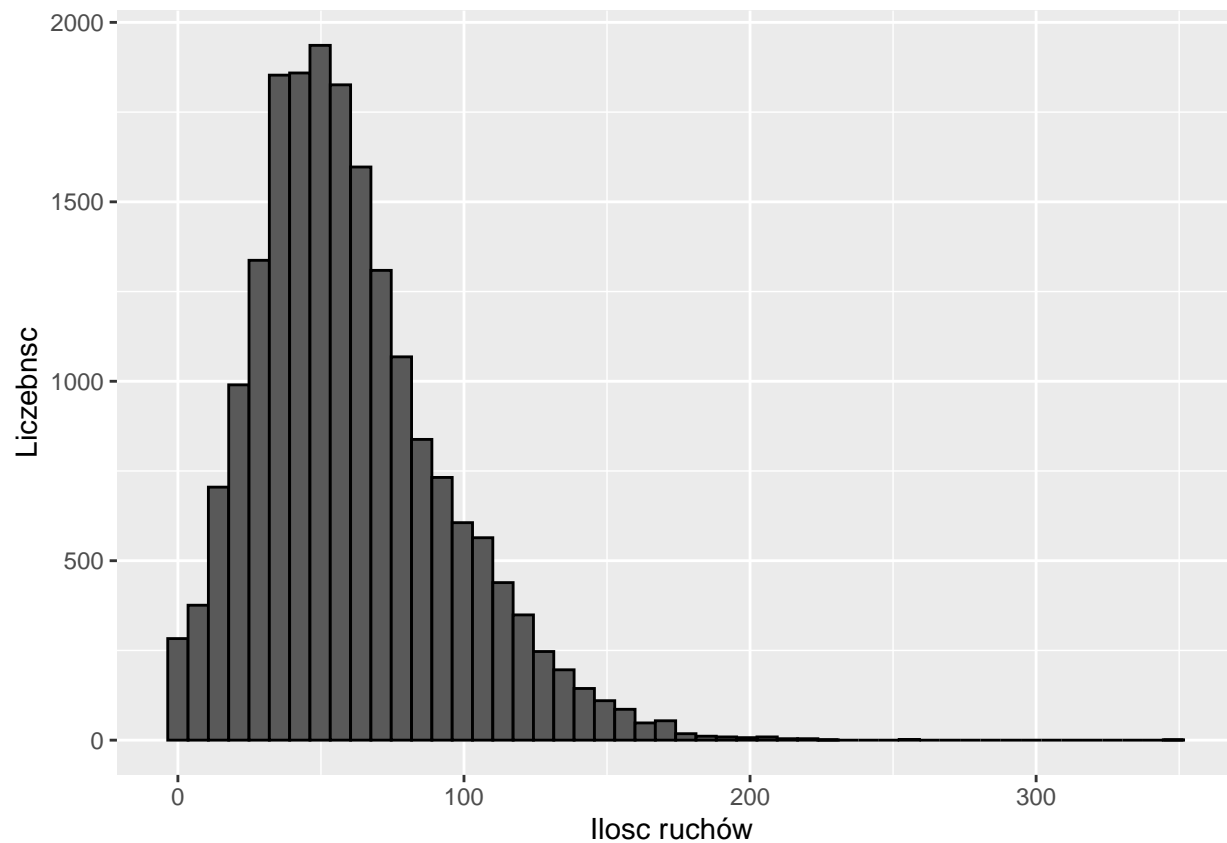


przedziałów które zawierają co najmniej 10 gier.

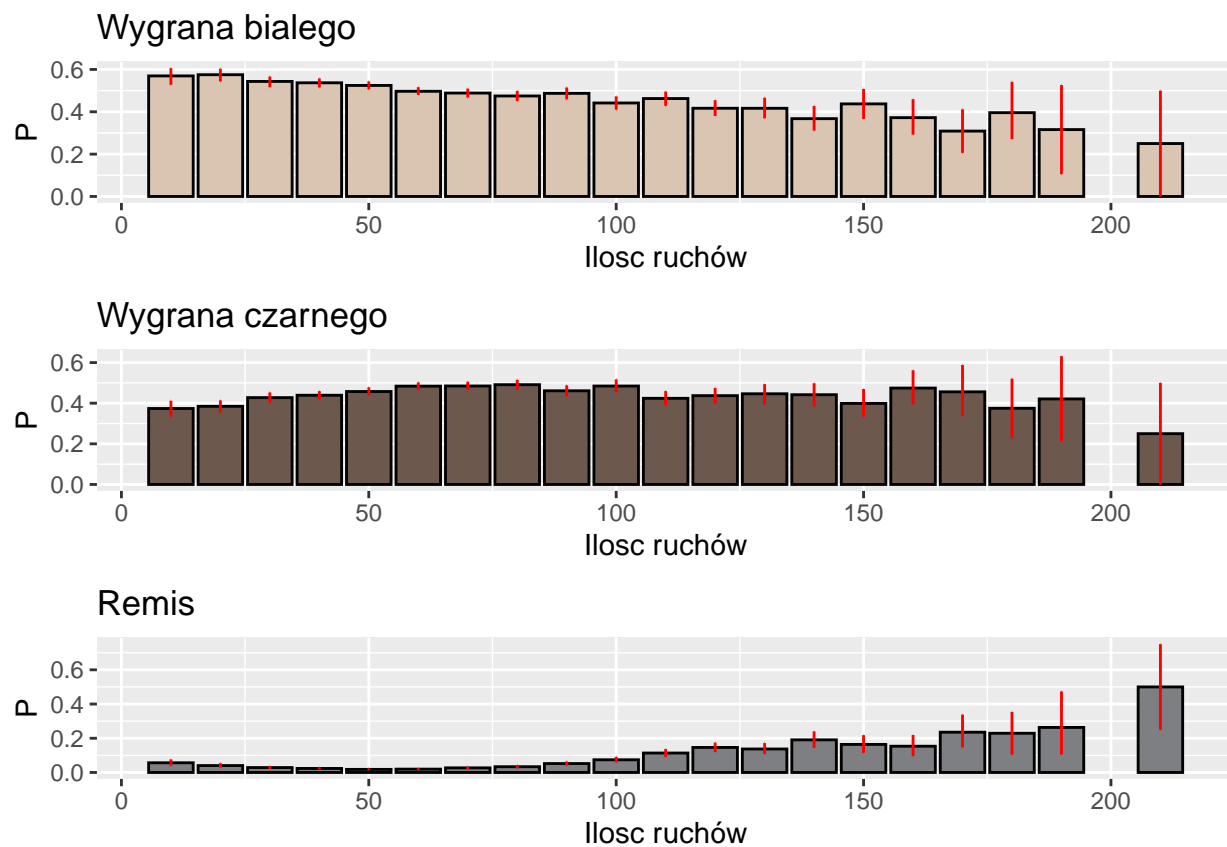
w powyższych wykresów wynika, że wraz ze wzrostem różnicy w rankingu rośnie prawdopodobieństwo wygranej białego, natomiast maleje szansa na wygraną czarnego. W przypadku remisu ciężko wywnioskować zależność.

### Warunkowanie ilością ruchów

Jak wiemy partie w szachach mają różne długości. Sprawdźmy najpierw ile trwają partie w naszych grach.



jak możemy załważyć większość gier kończy się po 100 ruchach. Z powodu, że nasze dane są ciągłe podzielimy je na przedziały o długości 10 gdzie pierwszy przedział to  $[0, 10)$ . Sprawdzimy jak wyglądają prawdopodobieństwa zakończenia partii w zależności od przedziałów ilości posunięć dla przedziałów które zawierają co najmniej 10 gier



z wykresów możemy odczytać że prawdopodobieństwo wygrania białych maleje wraz z kolejnymi róchami, natomiast prawdopodobieństwo wygrania czarnych rośnie do około 100 ruchów by potem maleć. Natomiast prawdopodobieństwo remisu rośnie wraz z kolejnymi posunięciami.