

# 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, 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.

## Obróbka danych

### 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.

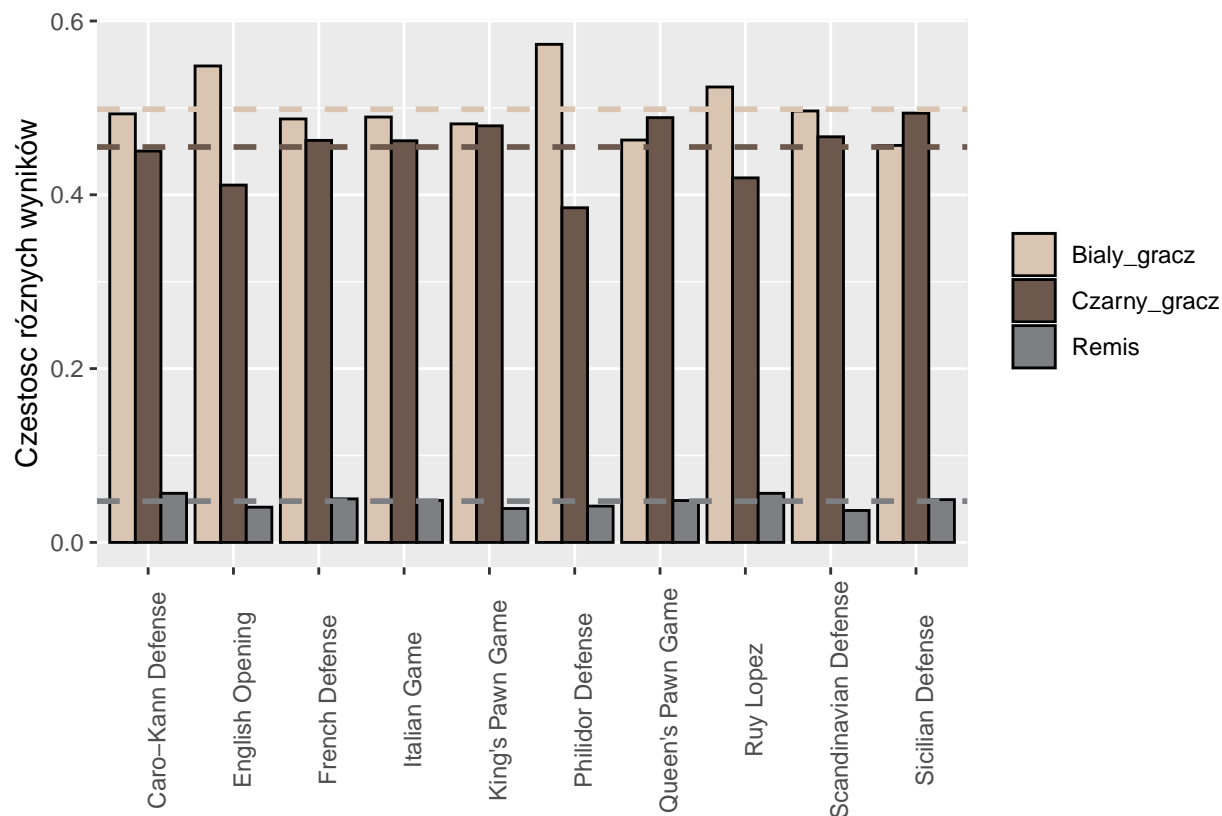
## Warunkowanie otwarciem

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:

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.  
## i Please use 'linewidth' instead.
```

```
##
```

```
## Scale for fill is already present.  
## Adding another scale for fill, which will replace the existing scale.  
##  
##  
## Scale for fill is already present.  
## Adding another scale for fill, which will replace the existing scale.
```



Przerywanymi 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.

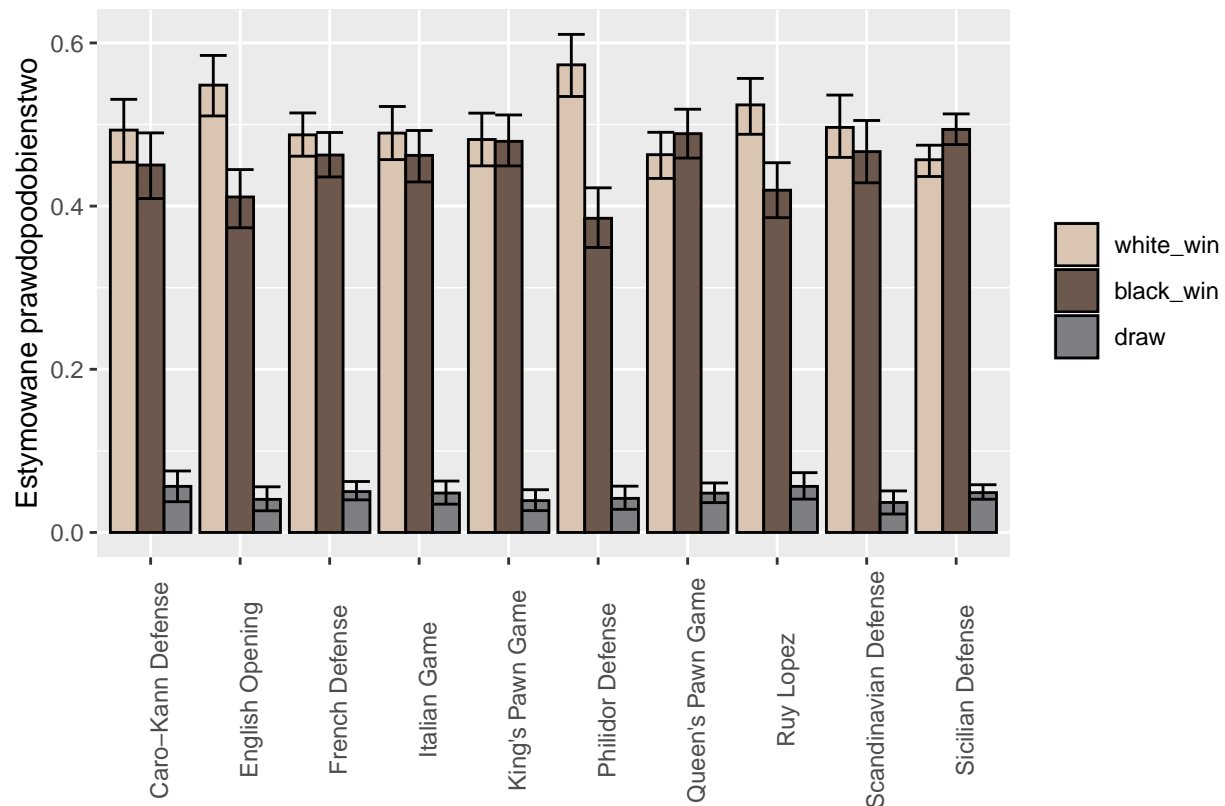
Przy dwóch z analizowanych otwarć gracz czarny nie wygrywał częściej niż gracz biały: przy obronie sycylijskiej ("Sicilian Defense") oraz gry pionkiem hetmańskim ("Queen's Pawn Game").

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

```
##
```

```
## Scale for fill is already present.
```

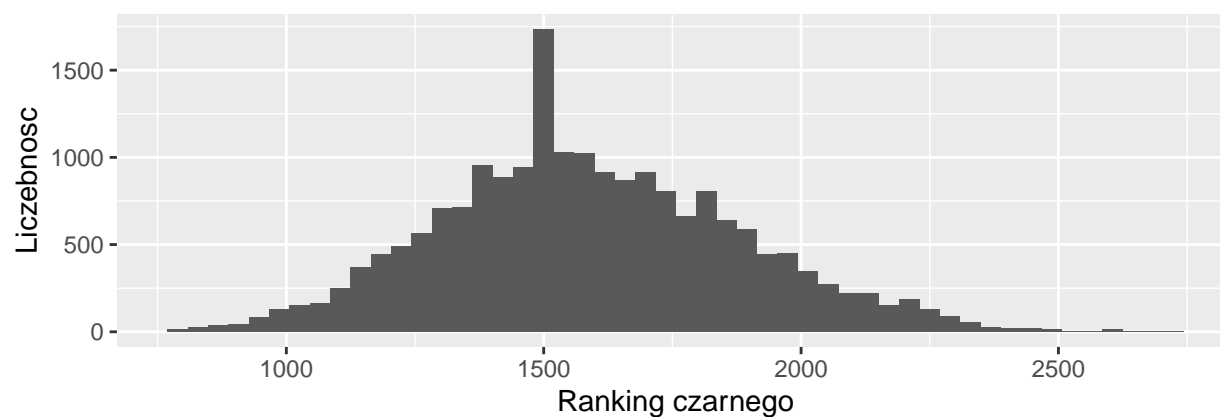
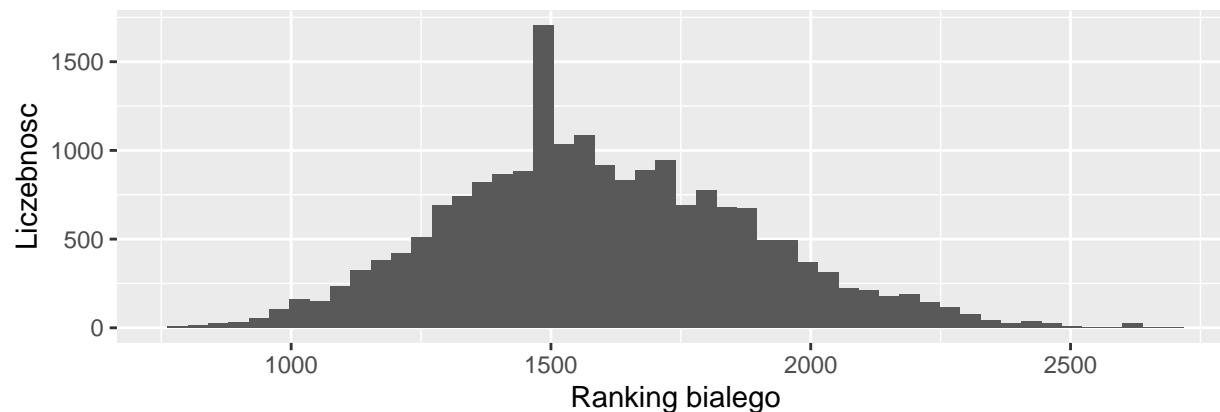
```
## Adding another scale for fill, which will replace the existing scale.
```



## Warunkowanie rankningiem białego

Jak wiemy gracze mają różne poziomy rankingi więc zobaczymy jak wyglądają prawdopodobieństwa zakończeń w zależności od rankingu białego. Napoczątku zobaczmy jak wygląda rozkład rankingu białego a jak czarnego.

```
plot_white_ratig <- ggplot(data=df, mapping=aes(x=white_rating)) + geom_histogram(bins=50)+
  labs(x = 'Ranking białego', y= 'Liczebność') # rozkład rankingu białego
plot_black_rating <- ggplot(data=df, mapping=aes(x=black_rating)) + geom_histogram(bins=50)+
  labs(x = 'Ranking czarnego', y="Liczebność") # rozkład rankingu czarnego
grid.arrange(plot_white_ratig,plot_black_rating)
```



Z wykresów możemy odczytać, że rozkład rankigu białego i rankingu czarnego są bardzo podobne. Teraz rodzi się pytanie jak wyglądają ich średnia, mediana, wariancja, skośność czy kurtოza

```
mean(df$white_rating)
```

```
## [1] 1596.146
```

```
mean(df$black_rating)
```

```
## [1] 1588.387
```

```
median(df$white_rating)
```

```
## [1] 1567
```

```
median(df$black_rating)
```

```
## [1] 1562
```

```
var(df$white_rating)
```

```
## [1] 84221.18
```

```
var(df$black_rating)
```

```
## [1] 84214.67
```

```
skewness(df$white_rating)
```

```
## [1] 0.2965193
```

```
skewness(df$black_rating)
```

```
## [1] 0.2518976
```

```
kurtosis(df$white_rating)
```

```
## [1] 0.02922972
```

```
kurtosis(df$black_rating)
```

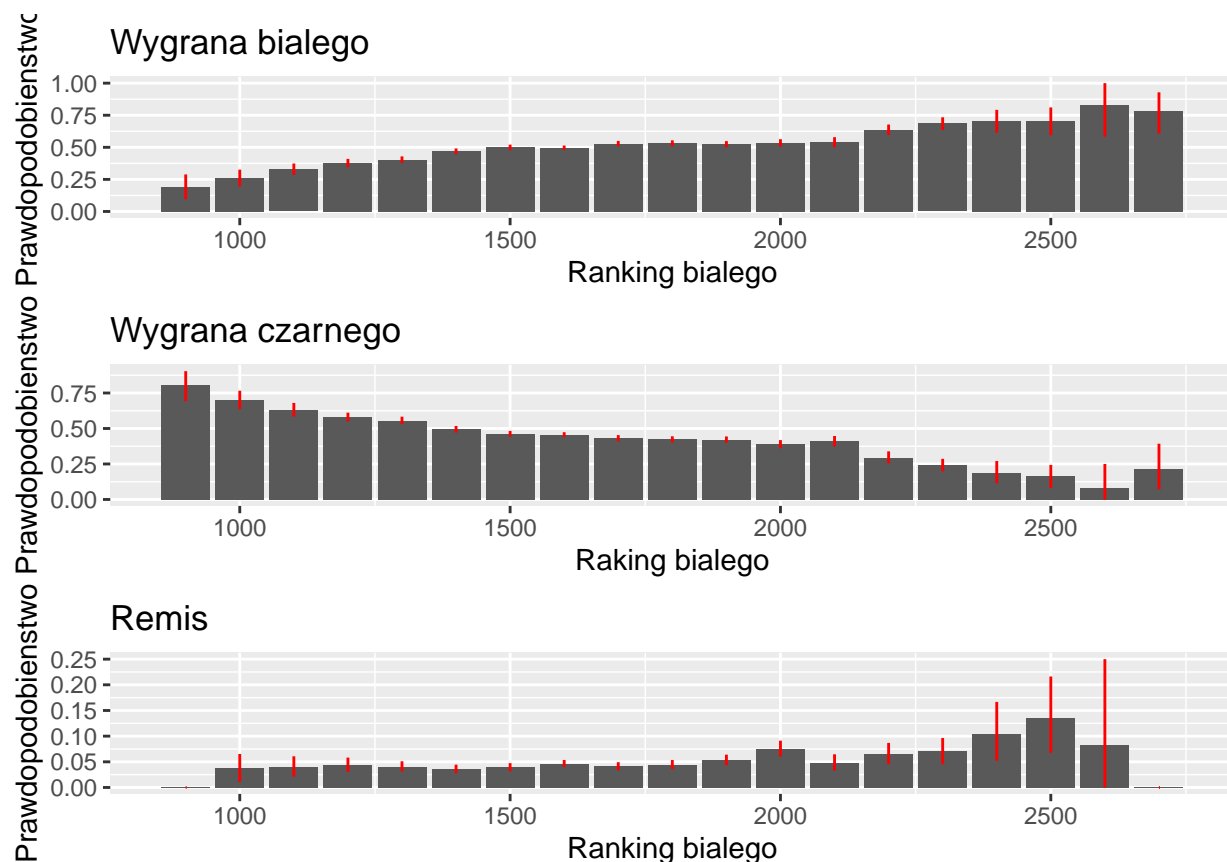
```
## [1] -0.052485
```

	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

Tak więc analizując statystyki opisowe możemy dojść do wniosku że biały ma średnio większy ranking ale tylko o około 8 punktów rankingowych co przy średnim rankingu nieco mniejszym niż 1600 jest znikomą przewagą, dodatkowo możemy odczytać że mediana białego 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 ranking jest zmienną ciągłą to podzielimy ranking białego na przedziały np [1300;1400) i taki przedział będziemy oznaczać 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 kreski pionowe to są przedziały ufności dla prawdopodobieństwa wygranej.

```
new_df = read.csv("winner_depend_ranking.csv")
plot_white <- ggplot(data=new_df,aes(x=values,y=white_win)) + geom_bar(stat='identity')+ geom_errorbar
  labs(x = 'Ranking białego', y = 'Prawdopodobieństwo', title="Wygrana białego")
plot_black <- ggplot(data=new_df,aes(x=values,y=black_win)) + geom_bar(stat='identity') + geom_errorbar
  labs(x = 'Ranking białego', y = 'Prawdopodobieństwo', title="Wygrana czarnego")
plot_draw <- ggplot(data=new_df,aes(x=values,y=draw)) + geom_bar(stat='identity')+ geom_errorbar(aes(ymin=
  labs(x = 'Ranking białego', y = 'Prawdopodobieństwo', title="Remis")
grid.arrange(plot_white,plot_black,plot_draw)
```



Analizując wykres możemy załważyć, że wraz 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 ufności 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 partii od rankingu białego, ale ranking czarnego powinien mieć też wpływ więc Dodajmy nową kolumnę, która będzie oznaczać różnice między rankingiem białego i czarnego. Różnice w rankingach ponieważ są ciągłe więc też podzielimy je na kategorię tak jak dla rankingów białego. Rozkład różnic w rankingach wygląda następująco:

```
ggplot(data=df,aes(x=ranking_dif)) + geom_histogram(bins=50) + labs(x= 'Różnica w rankingach', y='Li')
```

