

Analiza parametrów skoków w zawodach skoków narciarskich

406949, Dorota Meszka, Środa 12⁵⁰

*AGH, Wydział Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2020/2021*

Kraków, 27 stycznia 2022

Ja, niżej podpisany(na) własnoręcznym podpisem deklaruję, że przygotowanym(tam) przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....

1 Streszczenie raportu

Raport powstał, w oparciu o analizę danych dotyczących statystyk poszczególnych skoków w konkursach skoków narciarskich mężczyzn na skoczniach dużych w latach 2009-2022.

2 Opis danych

Dane do projektu pochodzą ze strony [Kaggle](#). Są one przedstawione w pliku .csv (plik all_results). Dane składają się z 288012 rekordów, każdy z nich zawiera 20 cech. Każdy rekord opisuje jeden skok, cechy oznaczają kolejno:

- **points** - łączna punktacja za skok (*numeryczna*)
- **speed** - szybkość zawodnika na progu (*numeryczna*)
- **dist** - odległość zawodnika (*numeryczna*)
- **dist_points** - punkty za odległość (*numeryczna*)
- **note_1** - pierwsza nota za styl (*numeryczna*)
- **note_2** - druga nota za styl (*numeryczna*)
- **note_3** - trzecia nota za styl (*numeryczna*)
- **note_4** - czwarta nota za styl (*numeryczna*)

- **note_5** - piąta nota za styl (*numeryczna*)
- **note_points** - łączna nota za styl (*numeryczna*)
- **gate** - numer belki startowej (*numeryczna*)
- **id** - ID konkursu (*kategoryczna*)
- **loc** - pozycja zawodnika (w punktacji końcowej serii/konkursu) (*numeryczna*)
- **bib** - numer startowy zawodnika (*kategoryczna*)
- **round** - określenie rundy (*kategoryczna*)
- **wind** - prędkość wiatru (w m/s) (*numeryczna*)
- **wind_comp** - wartość rekompensaty za wiatr (*numeryczna*)
- **gate_points** - punkty za zmianę rozbiegu (*numeryczna*)
- **Unnamed..0** - numer skoku (wg. wewnętrznej numeracji zawodów) (*numeryczna*)
- **codex** - FIS ID zawodnika (*numeryczna*)

Dane, z których korzystam, są w sporej części niekompletne:

```
> colSums(is.na(results))
```

	points	speed	dist	dist_points	note_1	note_2
	133729	4	4	28592	133729	133729
	note_3	note_4	note_5	note_points	gate	id
	133729	133729	133729	133729	28284	0
	loc	bib	round	wind	wind_comp	gate_points
	36691	0	0	58752	58752	68113
	Unnamed..0	codex				
	223364	757				

dlatego przed analizą przystąpię do ich czyszczenia.

3 Czyszczenie danych

Ze względu na obszerny rozmiar mojej bazy danych (prawie 300 tysięcy pozycji), usunięcie nawet sporej części danych zostawia wystarczająco dużą bazę, dlatego decyduję się na następujące kroki:

1. Usunięcie kolumny *Unnamed..0* - znaczna część jej komórek ma wartości nieznane a te, które nieznane nie są, i tak są dość bezużyteczne - numerowanie skoków w konkursie nie odbywa się bowiem na podstawie kolejności wykonanych skoków, a według punktacji końcowej, co wydaje mi się być nielogiczne

2. Usunięcie rzędów z pozostałych kolumn z wartościami nieznanymi - co prawda możnaby próbować wypełnić te dane wartościami średnimi, jednak musimy brać pod uwagę, że każde zawody/skocznia cechują się odpowiednimi parametrami i coś takiego mogłoby doprowadzić do uzyskania danych, które nie mają logicznego sensu (np. wysoki numer belki startowej której na skoczni może nie być)
3. Usunięcie rzędów, dla których wartość cechy *speed* wynosi 0 - tego typu anomalia może wynikać jedynie z błędów pomiarowych i rozsądny wydaje się pominięcie jej w analizie statystycznej.

Czyszczenie danych wykonuję za pomocą funkcji:

```
> results = subset(results, select =-c(Unnamed..0))
> row.has.na <- apply(results, 1, function(x){any(is.na(x))})
> results <- results[!row.has.na,]
> results <- results[!(results$speed == 0),]
```

Innym problemem, który utrudnia zaobserwowanie korelacji pomiędzy danymi jest wspólne zestawienie wyników ze skoczni normalnych, dużych oraz mamucich, oraz zawodów rozgrywanych dla płci męskiej i żeńskiej. Aby przefiltrować odpowiednio te dane, korzystam z danych z pliku *all_comps* (zapisanych jako *competitions*) i wprowadzam założenia, które eliminują te problemy:

```
> competitions <- competitions[!(competitions$gender == "Women" |
+   competitions$hill_size_y < 120 | competitions$hill_size_y >
+   150), ]
> results <- results[results$id %in% competitions$id, ]
```

Dane prezentują się teraz następująco:

```
> summary(results)
```

	points	speed	dist	dist_points	
	Min. : 0.0	Min. :82.30	Min. : 20.0	Min. :-120.00	
	1st Qu.: 97.6	1st Qu.:89.00	1st Qu.:116.0	1st Qu.: 49.20	
Median	:112.2	Median :90.60	Median :123.0	Median : 60.90	
Mean	:109.1	Mean :90.55	Mean :121.4	Mean : 58.19	
3rd Qu.	:123.9	3rd Qu.:92.00	3rd Qu.:128.5	3rd Qu.: 69.90	
Max.	:172.8	Max. :98.90	Max. :152.0	Max. : 114.00	
	note_1	note_2	note_3	note_4	note_5
	Min. : 3.0	Min. : 3.0	Min. : 3.0	Min. : 3.00	Min. : 3.0
	1st Qu.:17.0	1st Qu.:17.0	1st Qu.:17.0	1st Qu.:17.00	1st Qu.:17.0
Median	:17.5	Median :17.5	Median :17.5	Median :17.50	Median :17.5
Mean	:17.3	Mean :17.3	Mean :17.3	Mean :17.29	Mean :17.3
3rd Qu.	:18.0	3rd Qu.:18.0	3rd Qu.:18.0	3rd Qu.:18.00	3rd Qu.:18.0
Max.	:20.0	Max. :20.0	Max. :20.0	Max. :20.00	Max. :20.0
	note_points	gate	id		loc

```

Min.   : 9.00   Min.   : 0.00   Length:47213      Min.   : 1
1st Qu.:50.50  1st Qu.:10.00  Class :character  1st Qu.: 8
Median :52.50  Median :14.00  Mode  :character  Median :19
Mean   :51.92  Mean   :14.01
3rd Qu.:54.00  3rd Qu.:18.00
Max.   :60.00  Max.   :37.00

           bib          round         wind        wind_comp
Length:47213      Length:47213      Min.   :-3.4400    Min.   :-49.700
Class :character   Class :character  1st Qu.:-0.3500  1st Qu.:-6.700
Mode  :character   Mode  :character  Median : 0.1000  Median : -1.000
                           Mean   : 0.1686  Mean   : -1.312
                           3rd Qu.: 0.6600  3rd Qu.: 4.100
                           Max.   : 4.8600  Max.   : 43.500

gate_points       codex
Min.   :-33.2000  Min.   : 1905
1st Qu.: 0.0000  1st Qu.: 4664
Median : 0.0000  Median : 5564
Mean   : 0.2631  Mean   : 5555
3rd Qu.: 0.0000  3rd Qu.: 6551
Max.   : 48.2000 Max.   :219775

> dim(results)
[1] 47213     19

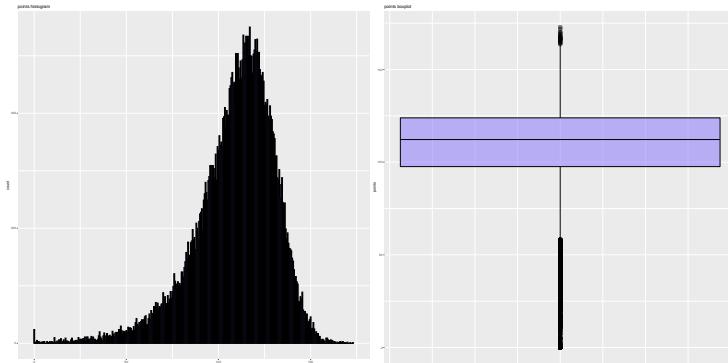
```

4 Analiza danych

Analizie danych nie poddaje się cech, które mają wartości kategoryczne: *bib*, *round*, *id*. Cechy o wartościach numerycznych *codex* oraz *loc* również są pomijane przede mnie w analizie - *codex* stanowi rodzaj identyfikatora, zatem jego wartość jest losowa, a *loc* jest wielokrotnością przedziałów 1-50, co jest nieszczególnie interesującym zbiorem danych do analizy.

Cechy *note_1*, *note_2*, *note_3*, *note_4* oraz *note_5* grupuję ze sobą w nową cechę, określoną przez *s_notes_data* - decyduję się na tego typu uproszczenie ze względu na losowe przypisanie tych danych do konkretnych cech - sędziowie nie są stali, zatem bardziej interesują nas pojedyncze noty jako ogół, aniżeli ich pogrupowanie.

4.1 Cecha *points*

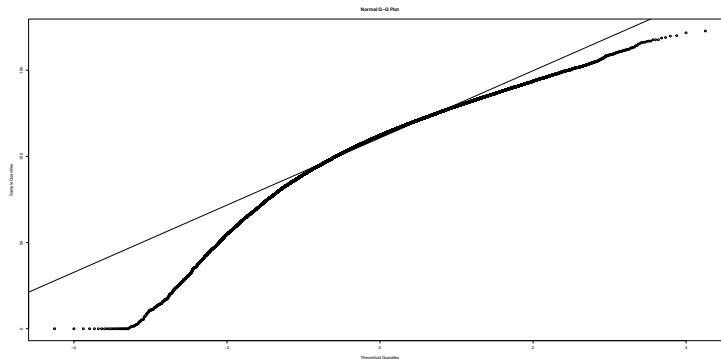


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
points 0 172.8 109.0661 476.2582 21.82334

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
points -1.001832 4.890825 97.6 112.2 123.9

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
points 108.8692 109.2629 470.2406 482.3925

Wartości znajdują się w przedziale od 0 do 172.8, ich średnia wynosi 109.0661, wariancja 476.2582, a odchylenie standardowe 21.82334. Skośność przyjmuje wartość ujemną, rozkład ten jest więc lewostronnie skośny. Kurtoza przyjmuje wartość >3 (kurtoza rozkładu normalnego), zatem układ jest leptokurtyczny.

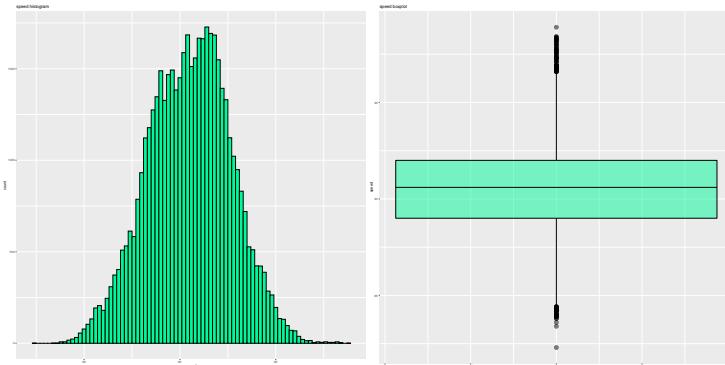


Anderson-Darling normality test

```
data: results$points
A = 468.97, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *points* nie jest rozkładem normalnym.

4.2 Cecha *speed*

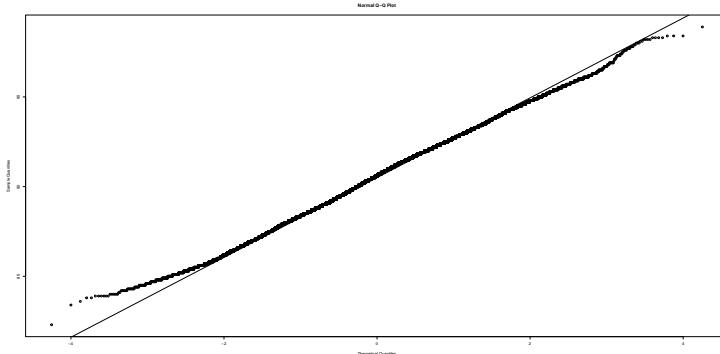


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
 speed 82.3 98.9 90.55327 4.643792 2.154946

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
 speed -0.04902606 2.802782 89 90.6 92

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
 speed 90.53383 90.57271 4.585117 4.703605

Cecha przyjmuje wartości pomiędzy 82.3 a 98.9, średnia wynosi 90.55327, wariancja 4.643792, a odchylenie standarowe 2.154946. Skośność przyjmuje wartość ujemną, rozkład ten jest więc lewostronnie skośny. Kurtoza przyjmuje wartość < 3, zatem układ jest platykurtyczny.

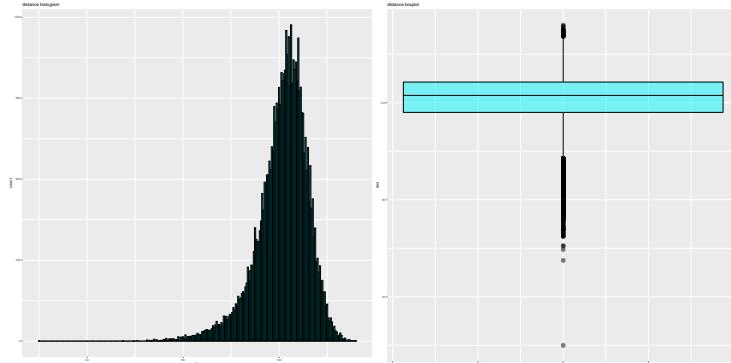


Anderson-Darling normality test

```
data: results$speed
A = 23.837, p-value < 2.2e-16
```

Mimo kurtozy o wartości zbliżonej do wartości kurtozy rozkładu normalnego oraz dzwonowego kształtu histogramu, wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *speed* nie jest rozkładem normalnym.

4.3 Cecha *dist*

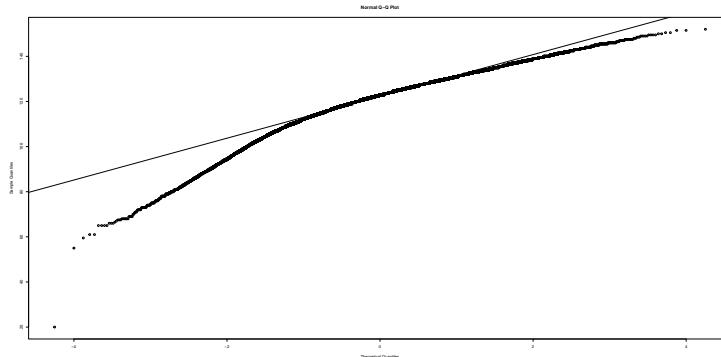


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
 dist 20 152 121.4144 112.3413 10.59912

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
 dist -0.9794381 5.042723 116 123 128.5

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
 dist 121.3187 121.51 110.9218 113.7883

Wartości zawierają się pomiędzy 20 a 152 [m], średnia wartość wynosi 121.4144 [m], wariancja 112.3413 [m], a odchylenie standardowe 10.59912 [m]. Skośność przyjmuje wartość ujemną, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza przyjmuje wartość > 3 , zatem układ jest leptokurytyczny.

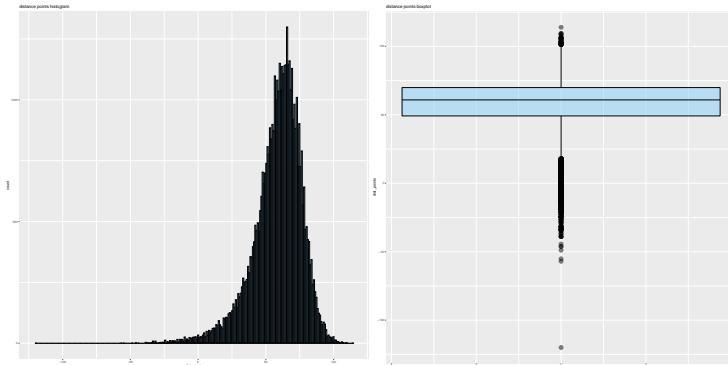


Anderson-Darling normality test

```
data: results$dist
A = 448.19, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *dist* nie jest rozkładem normalnym.

4.4 Cecha *dist_points*

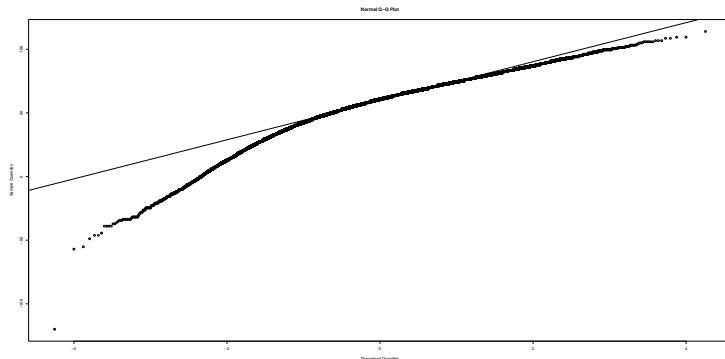


Zmienna Minimum Maksimum Średnia Wariancja Odch._stand.
dist_points -120 114 58.18943 323.4025 17.9834

Zmienna Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
dist_points -1.045896 5.32277 49.2 60.9 69.9

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
dist_points 58.02721 58.35165 319.3163 327.568

Cecha przybiera wartości między -120 a 114, średnia wartość wynosi 58.18943, wariancja 323.4025, a odchylenie standardowe 17.9834. Skośność przyjmuje wartość ujemną, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza > 3, zatem układ jest leptokurytyczny.

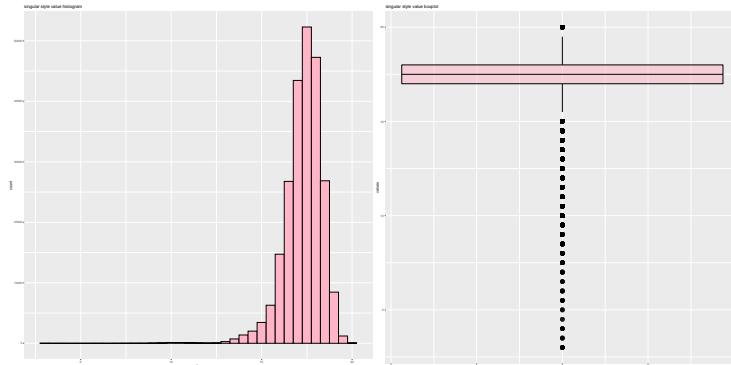


Anderson-Darling normality test

```
data: results$dist_points
A = 503.43, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *dist_points* nie jest rozkładem normalnym.

4.5 Cechy *note_1..5*

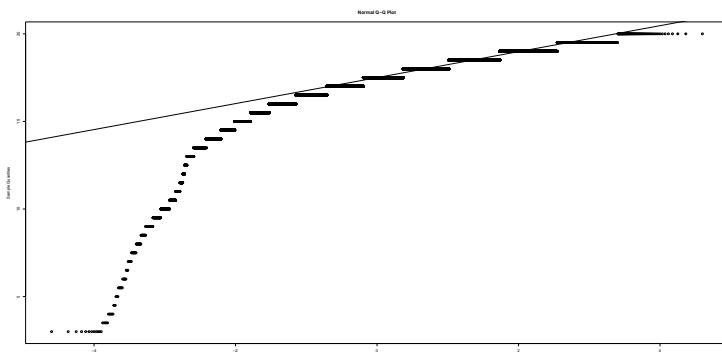


Cechy Minimum Maksimum Średnia Wariancja Odch._stand.
note_1..5 3 20 17.29987 1.155248 1.074825

Cechy Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
note_1..5 -1.88929 14.18501 17 17.5 18

Cechy CI_mean_start CI_mean_end CI_var_start CI_var_end
note_1..5 17.29553 17.3042 1.148686 1.161867

Wartości not zawierają się między 3 a 20 (można zaobserwować zgodność z regulami ocen sędziowskich), średnia wartość wynosi 17.29987, wariancja 1.155248, a odchylenie standardowe 1.074825. Skośność < 0, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza przyjmuje wartość > 3, zatem układ jest leptokurytyczny.

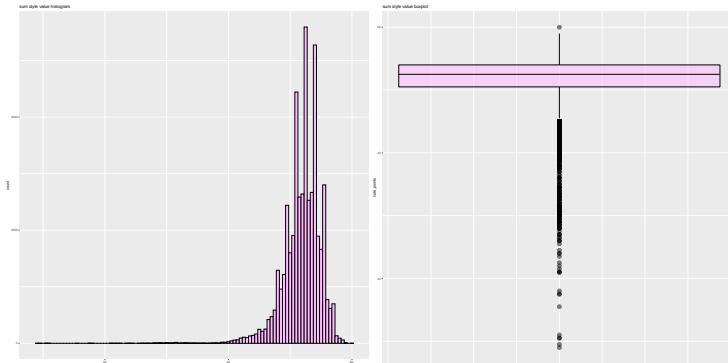


Anderson-Darling normality test

```
data: s_notes_data$values
A = 5019.7, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cech *note_1..5* nie jest rozkładem normalnym.

4.6 Cecha *note_points*

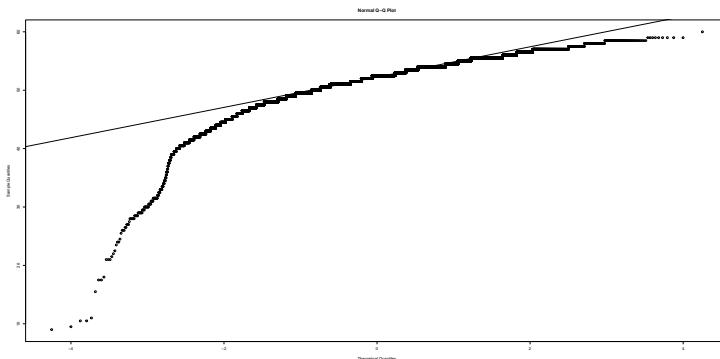


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
note_points 9 60 51.91537 9.580946 3.09531

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
note_points -2.04036 15.70238 50.5 52.5 54

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
note_points 51.88745 51.94329 9.45989 9.704351

Wartości sumy not zawierają się między 9 a 60 (kolejna zgodność z regułami oceniania FIS), średnia wartość wynosi 51.91537, wariancja 9.580946, a odchylenie standardowe 3.09531. Skośność przyjmuje wartość ujemną, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza przyjmuje wartość > 3 , zatem układ jest leptokurtyczny.

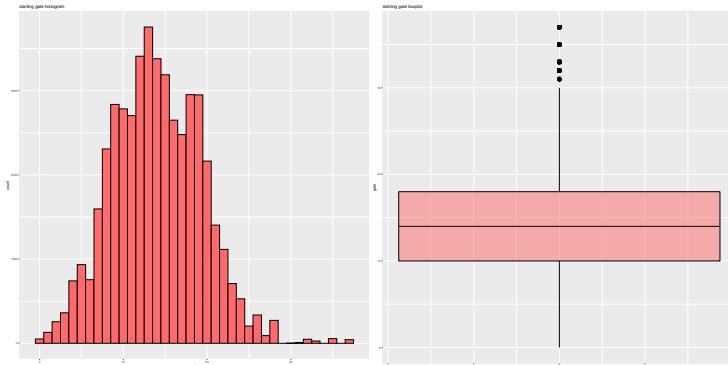


Anderson-Darling normality test

```
data: results$note_points
A = 674.2, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *note_points* nie jest rozkładem normalnym.

4.7 Cecha *gate*

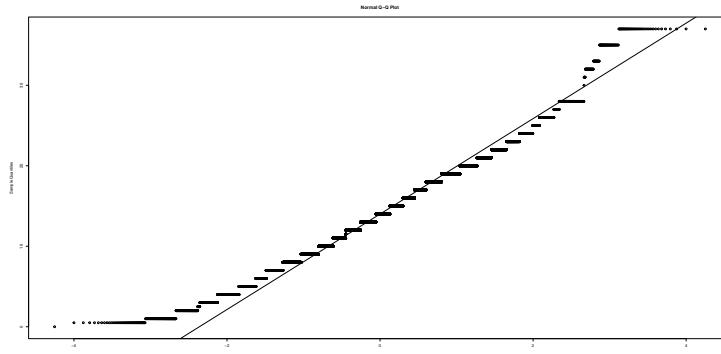


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
 gate 0 37 14.01133 28.1666 5.307221

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
 gate 0.2391563 3.219724 10 14 18

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
 gate 13.96346 14.05921 27.81071 28.52939

Cecha posiada wartości zawierające się między 0 a 37, średnia wartość wynosi 14.01133, wariancja 28.1666, a odchylenie standardowe 5.307221. Skośność przyjmuje wartość dodatnią, z czego wynika, że rozkład ten jest prawostronnie skośny. Kurtoza przyjmuje wartość > 3 , zatem układ jest leptokurtyczny.

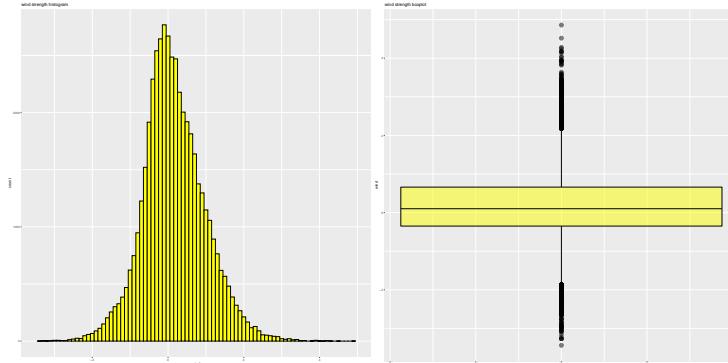


Anderson-Darling normality test

```
data: results$gate
A = 103.81, p-value < 2.2e-16
```

Mimo kurtozy o wartości zbliżonej do wartości kurtozy rozkładu normalnego oraz dzwonowego kształtu histogramu, wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *gate* nie jest rozkładem normalnym.

4.8 Cecha *wind*

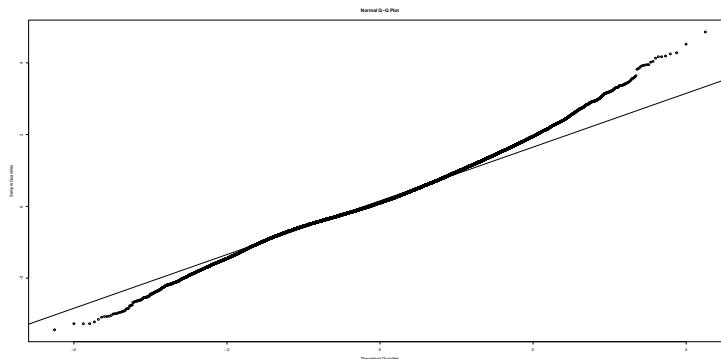


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
 wind -3.44 4.86 0.1686504 0.667926 0.8172674

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
 wind 0.2942379 3.881884 -0.35 0.1 0.66

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
 wind 0.1612782 0.1760225 0.6594868 0.6765291

Cecha posiada wartości zawierające się między -3.44 a 4.86 [m/s], średnia wartość wynosi 0.1686504 [m/s], wariancja 0.667926 [m/s], a odchylenie standarde 0.8172674 [m/s]. Skośność przyjmuje wartość dodatnią, z czego wynika, że rozkład ten jest prawostronnie skośny. Kurtoza przyjmuje wartość > 3 , zatem układ jest leptokuryczny.

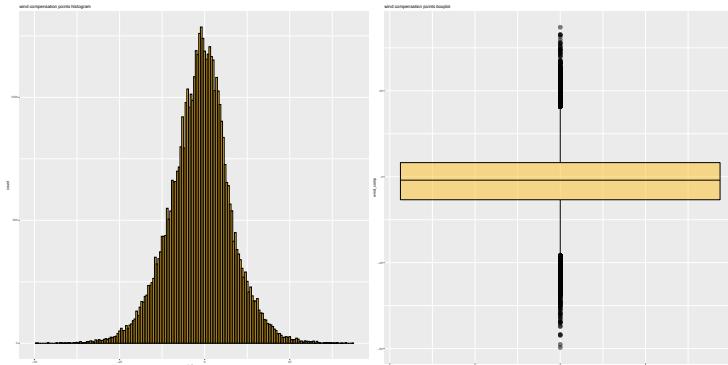


Anderson-Darling normality test

```
data: results$wind
A = 142.82, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *wind* nie jest rozkładem normalnym.

4.9 Cecha *wind_comp*

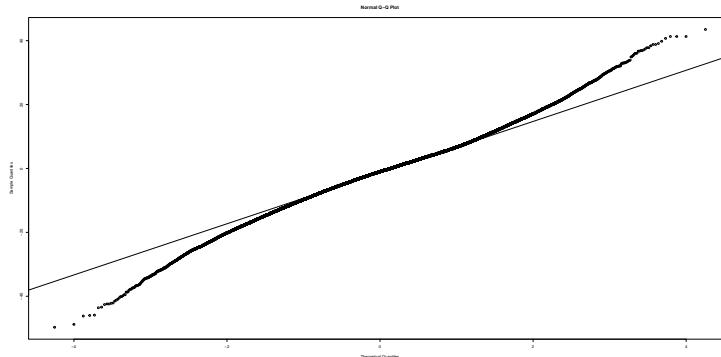


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
wind_comp -49.7 43.5 -1.311904 80.30611 8.961368

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
wind_comp -0.08047566 4.022278 -6.7 -1 4.1

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
wind_comp -1.392739 -1.231068 79.29144 81.34048

Wartości przeliczników za wiatr zawierają się w przedziale od -49.7 do 43.5, średnia wartość wynosi -1.311904, wariancja 80.30611, a odchylenie standardowe 8.961368. Skośność przyjmuje wartość ujemną, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza > 3, zatem układ jest leptokuryczny.

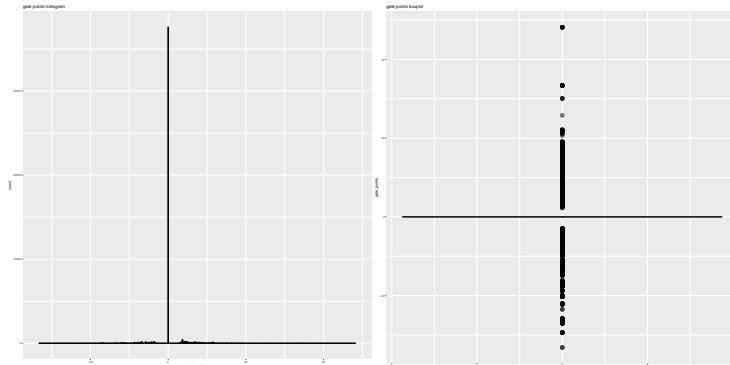


Anderson-Darling normality test

```
data: results$wind_comp
A = 92.818, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *wind_comp* nie jest rozkładem normalnym.

4.10 Cecha *gate_points*

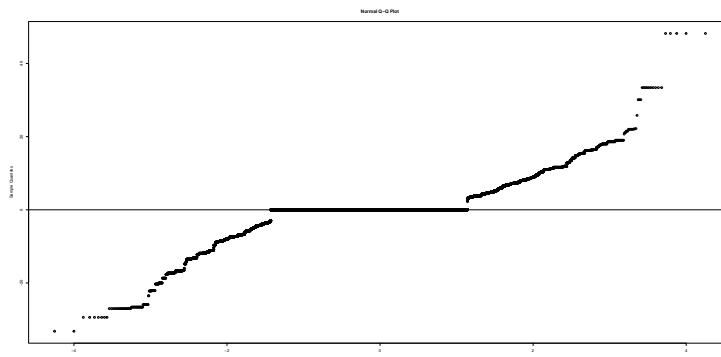


Cecha Minimum Maksimum Średnia Wariancja Odch._stand.
`gate_points` -33.2 48.2 0.2630737 13.068 3.614969

Cecha Skośność Kurtoza Kw_pierwszy Mediana Kw_trzeci
`gate_points` -0.04109729 18.79931 0 0 0

Cecha CI_mean_start CI_mean_end CI_var_start CI_var_end
`gate_points` 0.2304651 0.2956824 12.90288 13.23632

Wartości zawierają się w przedziale od -33.2 do 48.2, średnia wartość wynosi 0.2630737, wariancja 13.068, a odchylenie standardowe 3.614969. Skośność przyjmuje wartość ujemną, z czego wynika, że rozkład ten jest lewostronnie skośny. Kurtoza przyjmuje wartość > 3 , zatem układ jest leptokurytyczny.



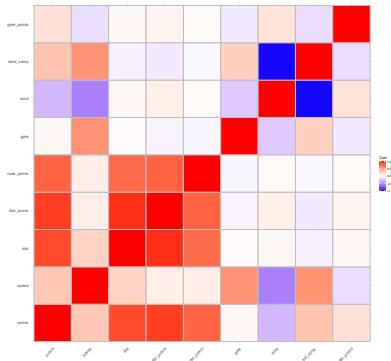
Anderson-Darling normality test

```
data: results$gate_points
A = 8848.1, p-value < 2.2e-16
```

Wykres QQ i test Andersona-Darlinga wskazują, że rozkład cechy *gate_points* nie jest rozkładem normalnym.

5 Analiza zależności między poszczególnymi cechami

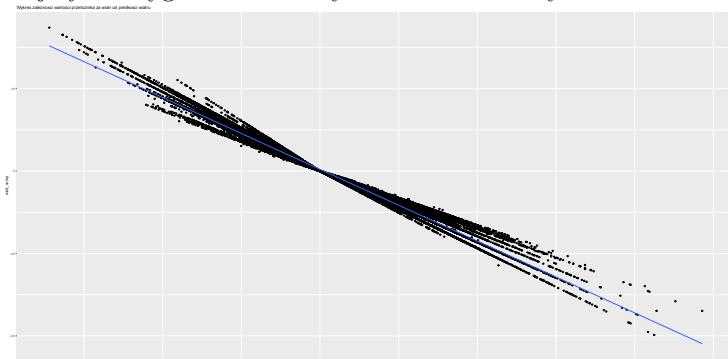
	points	speed	dist	dist_points	note_points
points	476.258150	13.6984578	197.9925355	353.923873	51.34117411
speed	13.698458	4.6437920	5.2276392	3.507518	0.58222583
dist	197.992535	5.2276392	112.3412681	178.662643	23.79680611
dist_points	353.923873	3.5075179	178.6626434	323.402525	42.67289099
note_points	51.341174	0.5822258	23.7968061	42.672891	9.58094600
gate	4.433529	6.4441625	1.2735707	-5.062547	-0.65759710
wind	-5.567788	-0.9705961	0.3713245	1.191298	0.06402488
wind_comp	59.698044	10.6897257	-5.3270510	-14.827078	-0.89118361
gate_points	12.367566	-1.0625004	1.5119713	3.816821	0.27979139
	gate	wind	wind_comp	gate_points	
points	4.4335294	-5.56778768	59.6980438	12.3675657	
speed	6.4441625	-0.97059609	10.6897257	-1.0625004	
dist	1.2735707	0.37132451	-5.3270510	1.5119713	
dist_points	-5.0625465	1.19129803	-14.8270785	3.8168209	
note_points	-0.6575971	0.06402488	-0.8911836	0.2797914	
gate	28.1665983	-1.01791394	11.9677025	-1.7979210	
wind	-1.0179139	0.66792604	-7.2690177	0.4367199	
wind_comp	11.9677025	-7.26901765	80.3061114	-4.7930868	
gate_points	-1.7979210	0.43671986	-4.7930868	13.0679984	
	points	speed	dist	dist_points	note_points
points	1.00000000	0.29128229	0.85596882	0.90181394	0.76004690
speed	0.29128229	1.00000000	0.22887564	0.09050901	0.08728727
dist	0.85596882	0.22887564	1.00000000	0.93732973	0.72534532
dist_points	0.90181394	0.09050901	0.93732973	1.00000000	0.76661298
note_points	0.76004690	0.08728727	0.72534532	0.76661298	1.00000000
gate	0.03827905	0.56345976	0.02264051	-0.05304325	-0.04003027
wind	-0.31217439	-0.55110959	0.04286667	0.08105588	0.02530931
wind_comp	0.30525627	0.55354876	-0.05608451	-0.09200461	-0.03212837
gate_points	0.15676838	-0.13639177	0.03946111	0.05871182	0.02500493
	gate	wind	wind_comp	gate_points	
points	0.03827905	-0.31217439	0.30525627	0.15676838	
speed	0.56345976	-0.55110959	0.55354876	-0.13639177	
dist	0.02264051	0.04286667	-0.05608451	0.03946111	
dist_points	-0.05304325	0.08105588	-0.09200461	0.05871182	
note_points	-0.04003027	0.02530931	-0.03212837	0.02500493	
gate	1.00000000	-0.23468194	0.25163398	-0.09371279	
wind	-0.23468194	1.00000000	-0.99251536	0.14782035	
wind_comp	0.25163398	-0.99251536	1.00000000	-0.14795732	
gate_points	-0.09371279	0.14782035	-0.14795732	1.00000000	



Analizując powyższe dane, w oczy rzuca się silna korelacja ujemna pomiędzy cechami *wind* oraz *wind_comp*. Innymi korelacjami, na które warto zwrócić uwagę, to te pomiędzy *speed* oraz *wind*, a także *note_points* oraz *dist*. Mimo widocznych korelacji między *note_points* oraz *points*, *dist_points* oraz *points* a również *points* i *dist*, decyduje się nie badać tych danych, ze względu na oczywistość ich korelacji - jako główne składowe/wartość przelicznika logicznym jest znaczny wpływ na wartość tej drugiej, a analizowana przeze mnie korelacja *wind* oraz *wind_comp* wystarczy na zaprezentowanie modelowego przykładu zgodności rozkładów.

5.1 Analiza zgodności *wind* i *wind_comp*

Zacznijmy od wygenerowania wykresu zależności tych dwóch cech:



Na podstawie wykresu widać silną zależność liniową, zatem można przystąpić do analizy regresji liniowej.

Call:

```
lm(formula = results$wind_comp ~ results$wind)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.9040	-0.5059	-0.3116	0.3880	9.9677

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.523513  0.005143 101.8   <2e-16 ***
results$wind -10.882968  0.006163 -1765.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

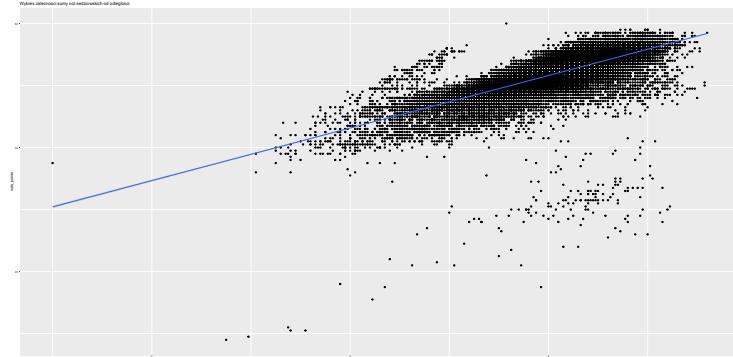
Residual standard error: 1.094 on 47211 degrees of freedom
Multiple R-squared:  0.9851,    Adjusted R-squared:  0.9851
F-statistic: 3.118e+06 on 1 and 47211 DF,  p-value: < 2.2e-16

```

Wartość współczynnika R^2 wynosi 0.9851, co świadczy o bardzo wysokim poziomie dopasowania. Współczynnik p-value potwierdza silną zależność między zmiennymi.

5.2 Analiza zgodności *note_points* i *dist*

Zacznijmy od wygenerowania wykresu zależności tych dwóch cech:



Na wykresie można zauważać zależność liniową, zatem przystępuję do analizy regresji liniowej.

```

Call:
lm(formula = results$note_points ~ results$dist)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.798  -0.798   0.172   1.113  10.185 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.620e+01  1.128e-01 232.3   <2e-16 ***
results$dist 2.118e-01  9.252e-04 228.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

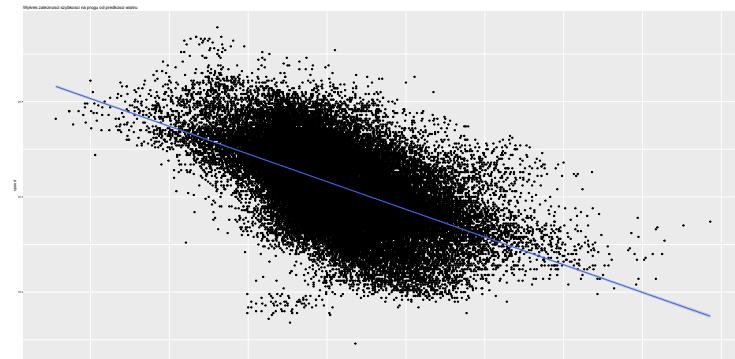
```

Residual standard error: 2.131 on 47211 degrees of freedom
Multiple R-squared:  0.5261,      Adjusted R-squared:  0.5261
F-statistic: 5.242e+04 on 1 and 47211 DF,  p-value: < 2.2e-16

```

Stosunkowo niska wartość $R^2 = 0.5261$ świadczy o nieszczególnie wysokim poziomie dopasowania. Mimo tego, wartość p-value sugeruje o silnej relacji między tymi cechami.

5.3 Analiza zgodności *speed* i *wind*



Zauważalna na wykresie jest zależność liniowa między cechami, zatem można przystąpić do analizy regresji liniowej.

```

Call:
lm(formula = results$speed ~ results$wind)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3660 -1.1425  0.0347  1.1735  7.1146 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 90.79834   0.00845 10745.4   <2e-16 ***
results$wind -1.45315   0.01013  -143.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 1.798 on 47211 degrees of freedom
Multiple R-squared:  0.3037,      Adjusted R-squared:  0.3037
F-statistic: 2.059e+04 on 1 and 47211 DF,  p-value: < 2.2e-16

```

Niska wartość współczynnika $R^2 = 0.3037$ świadczy o niewysokim poziomie dopasowania, natomiast niska wartość p-value sugeruje, że istnieje relacja między tymi cechami.

6 Wnioski

- Cechy *speed* oraz *gate* na pierwszy rzut oka mogłyby się wydawać, że należą do rozkładu normalnego, wątpliwości rozwiewa natomiast test normalności, który w przypadku obu cech zaprzecza tej hipotezie.
- Cechy *wind* oraz *wind_comp* są silnie zależne od siebie - jest to wniosek, którego należało się spodziewać, natomiast świadczy o poprawności danych oraz analizy. Dokładny sposób przeliczania wartości wiatru na punkty nie jest publicznie udostępniony, dzięki analizie danych można natomiast wnioskować, że musi być on zbliżony do liniowego.
- Występuje pewne, aczkolwiek nieszczególnie silne zjawisko korelacji między cechami *note_points* oraz *dist* - można zatem wnioskować, że sędziowie mają tendencję do lepszego punktowania skoków dłuższych, mimo że wedle reguł, długość skoku nie powinna wpływać na jego ocenę techniczną.
- Istnieje również korelacja między wartościami cechy *wind* oraz *speed* - choć niewielka, to pokazuje, że wiatr ma wpływ nie tylko na kształtowanie toru lotu zawodnika, ale wpływa też na niego na rozbiegu.
- Dziwić może natomiast niemalże całkowity brak korelacji między cechami *dist* oraz *wind* - sugeruje to, że kierunek i siła wiatru nie mają wpływu na długość oddawanych skoków, czemu zaprzeczą zarówno luźne obserwacje kibiców podczas zawodów, a także system rekompensaty punktowej za warunki wietrzne, który wprowadzony został przecież nie bez powodu. Być może jest to kwestia doboru danych (należałyby brać wyniki oscylujące tylko w określonym, węższym przedziale odległości), natomiast jest to dość zaskakująca obserwacja.