Analiza przebiegu półmaratonu, Wrocław 2023 metodą EDA

Autorzy:

Wnioski i komentarze – Dariusz Samól

Tabele i grafika: ChatGPT

Półmaraton, Wrocław 2023 - Analiza

- Nocny Półmaraton Wrocław. 8 197 zawodników punktualnie o godzinie 22 wyruszyło w trasę liczącą 21 km. Uczestnicy biegli m.in. przez plac Grunwaldzki, ulicami Pułaskiego, Kołłątaja, Podwale i Grodzką. Półmaraton ukończyło 8135 śmiałków.
  Źródło: https://www.tuwroclaw.com/wiadomosci,nocny-polmaraton-wroclaw-2023-tysiace-biegaczy-i-kibicow-na-trasie-nowe-zdjecia,wia5-3317-70799.html
- Zarejestrowane dane pokazały 8 150 miejsc świadczących o ukończeniu biegu, a więc 15 osób mogło być zdyskwalifikowanych (albo redakcja portalu nie miała kompletnych danych - hipoteza)
- 9% uczestników, którym przydzielono numery startowe nie ukończyło biegu (z porównania z informacją prasową wynika, że nie wszyscy stanęli do wyścigu).
- Zwycięzcą 9.Nocnego Wrocław Półmaratonu został Tomasz Grycko z czasem 1:04:59. Utrzymał on bardzo duże tempo biegu znacznie przewyższające tempo pozostałych uczestników.

```
ds.chat("Kto zwyciężył?")
```

```
'TOMASZ GRYCKO, Miejsce: 1.0, Czas: 01:04:59, Wiek: 31.0, Tempo: 3.0805088093545074, Tempo stabilność: 0.0313999999999999'
```

- Cała podstawowa grupa maratończyków utrzymała stabilne tempo w granicach ~ 4,8 …. ~ 6,7 min /km., lekko zwalniając z upływem kilometrów
- Zarówno mężczyźni, jak i kobiety pokonali półmaraton w większości stabilnym tempem.
- Oprócz zawodników z Polski w wydarzeniu wzięli udział zawodnicy z 20 krajów
- Najstarszy zawodnik miał 89 lat (rok temu, rocznik 1934 !) i zajął 8137 miejsce. Najmłodszy zawodnik mając 17 lat (rocznik 2006) ukończył bieg na 1604 pozycji.
- Z przeprowadzonej analizy wynika, że szanse ukończenia biegu na wyższych pozycjach mieli uczestnicy, którzy utrzymywali przez cały wyścig stabilne tempo.
- Kolejne kroki prezentują slajdy ułożone według schematu EXPLORATORY DATA ANALYSIS (EDA)

Eksploracyjna Analiza Danych (EDA)

Krok 1 Przegląd danych

Krok 2 Analiza brakujących wartości

Krok 3 Analiza zmiennych

Krok 4 Zastąp brakujące dane średnią

Krok 5 Analiza relacji pomiędzy zmiennymi

Krok 6 Analiza wartości odstających

Slajd 1/3 ~ Krok 1: Przegląd danych

Zrozumienie danych: jak były zbierane? Co oznaczają poszczególne wartości?

```
[55]: ds.chat("Pokaż ile jest rekordów")

      Please enter your OpenAI API token:  ·····
[55]: 8950

[4]: ds.chat("Pokaż 10 losowych rekordów")
```

[4]:

	Miejsce	Numer startowy	Imię	Nazwisko	Miasto	Kraj	Drużyna	Płeć	Płeć Miejsce	Kategoria wiekowa	Kategoria wiekowa Miejsce	Rocznik	5 km Czas	5 km Miejsce Open	5 km Tempo	
2160	2161.0	7702	KONRAD	KACZMAREK	WROCŁAW	POL	STAR NERD	M	1951.0	M30	695.0	1989.0	00:24:51	2168.0	4.970000	00
976	977.0	5563	AGNIESZKA	OSIŃSKA	GŁOGÓW	POL	KGHM ZG RUN ŚLĘZAK TEAM	K	68.0	K40	16.0	1980.0	00:22:42	876.0	4.540000	00
5408	5409.0	6385	AGNIESZKA	TORZEWSKA	WARSZAWA	POL	BEMOWO BIEGA	K	1065.0	K40	385.0	1974.0	00:29:27	5737.0	5.890000	00
3082	3083.0	3869	DARIUSZ	BONIECKI	SZCZECIN	POL	NaN	M	2695.0	M40	1004.0	1983.0	00:26:07	3127.0	5.223333	00
1429	1430.0	8739	PIOTR	WÓJCIK	TRZEBNICA	POL	NaN	M	1322.0	M40	480.0	1974.0	00:24:37	1989.0	4.923333	00
8832	NaN	6260	BARBARA	TUCZYNSKA	NaN	NaN	Active Team Bez Balastu	K	NaN	K40	NaN	1979.0	NaN	NaN	NaN	
4743	4744.0	939	RAFAŁ	KOSMALSKI	WROCŁAW	POL	NaN	M	3900.0	M50	396.0	1973.0	00:30:18	6235.0	6.060000	00
4619	4620.0	4555	MARCIN	BALICKI	WROCLAW	POL	NaN	M	3810.0	M20	662.0	2000.0	00:27:05	3924.0	5.416667	00
8053	8054.0	8094	PRZEMYSŁAW	KAWA	TARNOWSKIE GÓRY	POL	NaN	M	5780.0	M20	1003.0	1999.0	00:34:49	7860.0	6.963333	01
1404	1405.0	8475	ANTONI	KOŁAT	WROCŁAW	POL	NaN	M	1299.0	M20	213.0	2003.0	00:22:58	992.0	4.593333	00

Wnioski i komentarze

- W kolumnie „Miejsce” są brakujące wartości. Prawdopodobnie część uczestników nie dobiegła.
- Podobnie jest w kolumnach „Miasto” i „Kraj”.
- Tabela pozwala na wstępne zapoznanie się ze strukturą danych i typowymi wartościami komórek.

Slajd 2/3 ~ Krok 1: Przegląd danych

Zrozumienie danych: ile jest wartości unikatowych w każdej kolumnie?

```
[5]: ds.chat("Pokaż ile każda kolumna ma warto
```

[5]:

	Unique Values
Miejsce	8150
Numer startowy	8950
Imię	712
Nazwisko	6049
Miasto	1446
Kraj	31
Drużyna	1985
Płeć	2
Płeć Miejsce	5829
Kategoria wiekowa	13
Kategoria wiekowa Miejsce	1987
Rocznik	65
5 km Czas	1150
5 km Miejsce Open	8123
5 km Tempo	1150
10 km Czas	2109
10 km Miejsce Open	8139
10 km Tempo	1288
15 km Czas	2943
15 km Miejsce Open	8141
15 km Tempo	1402
20 km Czas	3703
Tempo Stabilność	4667
Czas	3800
Tempo	3798

Wnioski i komentarze

- Z brakujących danych wynika, że 800 osób nie ukończyło biegu (9% wszystkich uczestników). Można przypuszczać, że wstępną rejestrację uczestników prowadzono w chwili przydzielania numerów startowych, zaś pełnego zbierania danych o nich dokonywano już na mecie. Ci, którzy nie dobiegli nie przekazali danych.
- Obliczenie % populacji, która nie ukończyła biegu:
  - 100% * x/(y+x)
  - gdzie:
  - x: Ilość brakujących miejsc
  - y: Ilość miejsc zarejestrowanych

Slajd 3/3 ~Krok 1: Przegląd danych

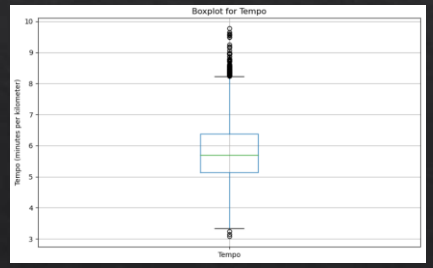
Zrozumienie danych: dla każdej kolumny – suma, wartość średnia, odchylenie standardowe, wartość minimalna, percentyle (pierwszy, drugi, trzeci, wartość maksymalna).

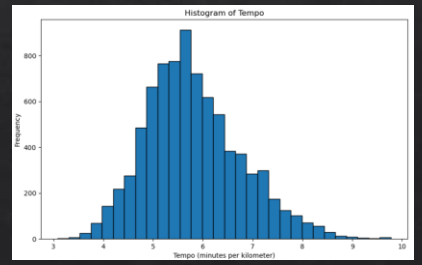
```
ds.chat('Pokaż podsumowanie każdej kolumny numerycznej').T
```

	count	mean	std	min	25%	50%	75%	max
Miejsce	8150.0	4075.500000	2352.846680	1.000000	2038.250000	4075.500000	6112.750000	8150.000000
Numer startowy	8950.0	4758.451732	2645.207126	1.000000	2504.250000	4770.500000	7010.750000	10000.000000
Płeć Miejsce	8150.0	2415.486626	1667.304693	1.000000	1019.250000	2038.000000	3791.750000	5829.000000
Kategoria wiekowa Miejsce	8141.0	649.992630	524.040520	1.000000	220.000000	517.000000	949.000000	1987.000000
Rocznik	8749.0	1980.938393	71.027734	0.000000	1977.000000	1984.000000	1991.000000	2006.000000
5 km Miejsce Open	8123.0	4070.677582	2350.132112	1.000000	2035.500000	4071.000000	6104.500000	8147.000000
5 km Tempo	8123.0	5.492411	0.807535	2.923333	4.936667	5.446667	6.016667	12.750000
10 km Miejsce Open	8139.0	4076.570586	2353.292768	1.000000	2038.500000	4076.000000	6113.500000	8156.000000
10 km Tempo	8116.0	5.536863	0.893716	2.926667	4.906667	5.456667	6.070833	9.753333
15 km Miejsce Open	8141.0	4075.617246	2352.375413	1.000000	2038.000000	4075.000000	6112.000000	8153.000000
15 km Tempo	8136.0	5.834662	0.999001	3.106667	5.142500	5.710000	6.410833	10.350000
20 km Miejsce Open	8144.0	4074.797151	2351.825153	1.000000	2038.750000	4074.500000	6111.250000	8148.000000
20 km Tempo	8137.0	6.481731	1.275844	3.386667	5.596667	6.260000	7.173333	14.940000
Tempo Stabilność	8110.0	0.065681	0.049127	-0.345333	0.033467	0.053067	0.085783	0.629533
Tempo	8150.0	5.806409	0.961578	3.080509	5.137868	5.690132	6.377499	9.775618

Wnioski i komentarze

- Analiza tempa
  - Średnie tempo: 5,8 min/km
  - Odchylenie standardowe: 0,96
  - Wartość min.: 3,1 min/km, wartość max.: 9,8 min/km
  - Przedział tempa: 4,8 – 6,7 min /km. Na poniższych wykresach widać duże skupisko biegaczy biegnących tempem w tym przedziale.
    Uwaga: Bloxplot pokazuje zakres w granicach 5,1-6,5





Slajd 1/1 ~ Krok 2: Analiza brakujących wartości

Ilość wartości brakujących i ich procentowy udział w całym zbiorze

```
[7]: ds.chat("Pokaż ile każda kolumna ma brakujących wartości oraz ile to stanowi procentowo")
```

[7]:

	Missing Values	Percentage
Miejsce	800	8.938547
Numer startowy	0	0.000000
Imię	0	0.000000
Nazwisko	0	0.000000
Miasto	1089	12.167598
Kraj	800	8.938547
Drużyna	5529	61.776536
Płeć	2	0.022346
Płeć Miejsce	800	8.938547
Kategoria wiekowa	11	0.122905
Kategoria wiekowa Miejsce	809	9.039106
Rocznik	201	2.245810
5 km Czas	827	9.240223
5 km Miejsce Open	827	9.240223
5 km Tempo	827	9.240223
10 km Czas	811	9.061453
10 km Miejsce Open	811	9.061453
10 km Tempo	834	9.318436
15 km Czas	809	9.039106
15 km Miejsce Open	809	9.039106
15 km Tempo	814	9.094972
20 km Czas	806	9.005587
20 km Miejsce Open	806	9.005587
20 km Tempo	813	9.083799
Tempo Stabilność	840	9.385475
Czas	0	0.000000
Tempo	800	8.938547

Wnioski i komentarze

- Brakujące wartości dla miasta i kraju potwierdzają hipotezę z pkt. **Slajd 2/3 ~ Krok 1: Przegląd danych**

```
20]: ds.chat("Dla każdej brakujące wartości 'Miejsce' pokaż Miasto' 'Kraj")
```

20]:

	Miasto	Kraj
8150	NaN	NaN
8151	NaN	NaN
8152	NaN	NaN
8153	NaN	NaN
8154	NaN	NaN
...	...	...
8945	NaN	NaN
8946	NaN	NaN
8947	NaN	NaN
8948	NaN	NaN
8949	NaN	NaN

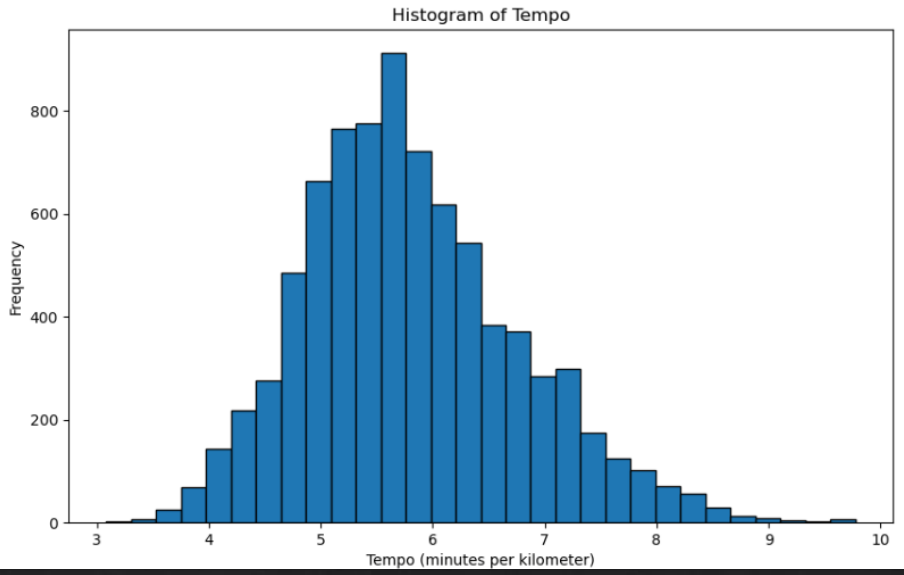
800 rows × 2 columns

Slajd 1/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: TEMPO

```
[8]: ds.chat("Narysuj histogram dla Tempa")
```

[8]:



Wnioski i komentarze

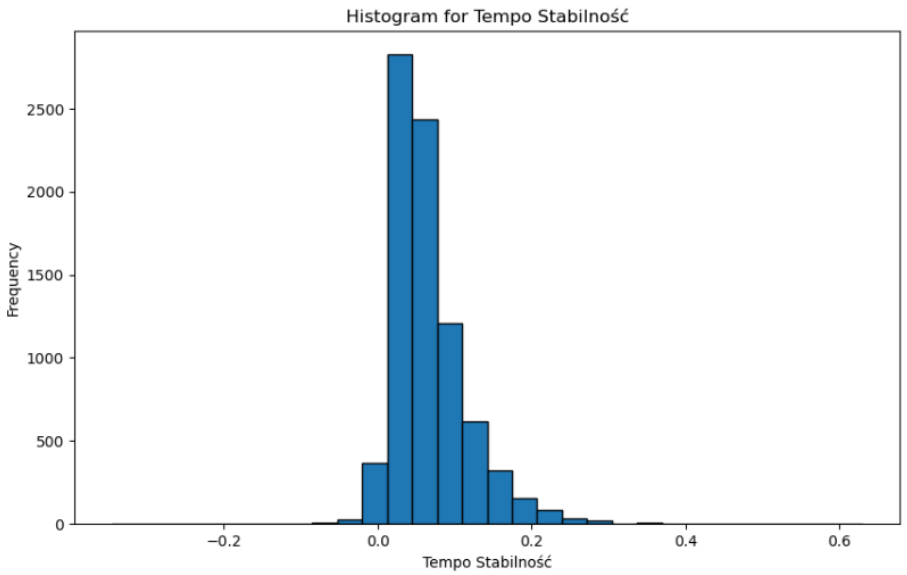
- Analiza tempa – ciąg dalszy
  - Wykres potwierdza obserwację z pkt. **Slajd 3/3 ~Krok 1: Przegląd danych**
  - Widać na nim duże skupisko biegaczy utrzymujących tempo w przedziale: ~ 4,8 …. ~ 6,7 min /km.

Slajd 2/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: STABILNOŚĆ

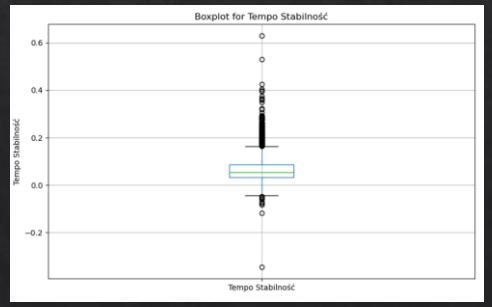
```
[9]: ds.chat("Narysuj histogram dla Tempo Stabilność")
```

[9]:



Wnioski i komentarze

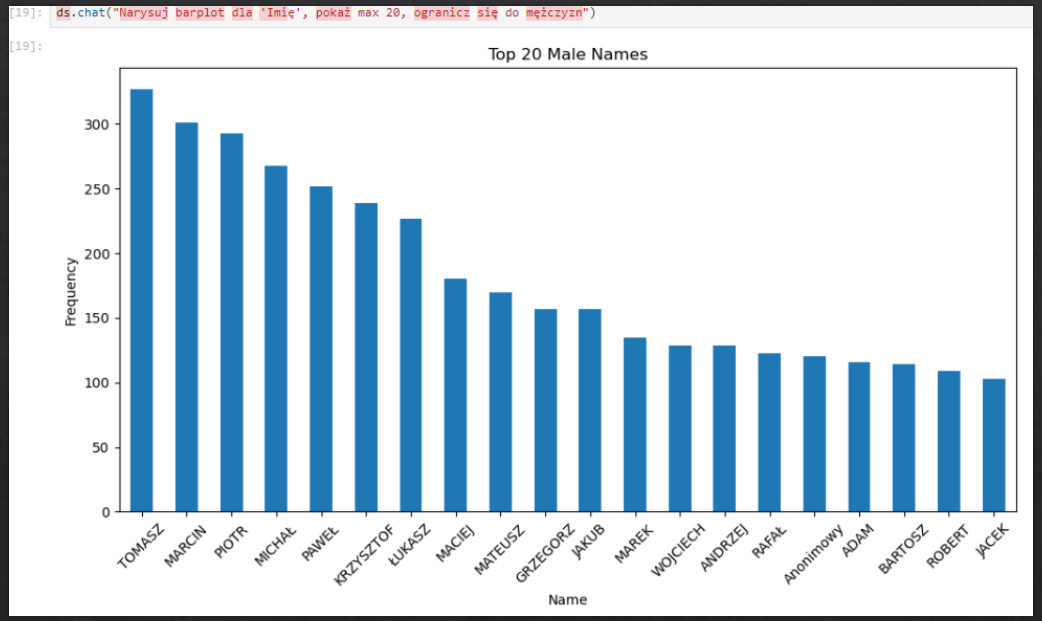
- Analiza tempa – ciąg dalszy
  - W porównaniu do półmaratonu następnego roku(2024) niewielka ilość biegaczy potrafiła trwale zwiększać tempo w czasie całego biegu. Naturalnym procesem uwarunkowanym fizjologicznie wydaje się zwalnianie spowodowane postępującym zmęczeniem.
  - Ilustruje to poniższy wykres. Widać na nim, że podstawowa grupa zawodników nieznacznie zwalniała. Kilku „outliers” potrafiło przyspieszyć (prawdopodobnie zawodowcy). Za tą grupą ciągnął się zmniejszający „ogon” tych, którzy walczyli, ale dobiegli (wiemy, że napewno dobiegli, bo mamy ich dane).



Slajd 3/9 ~ Krok 3: Analiza zmiennych

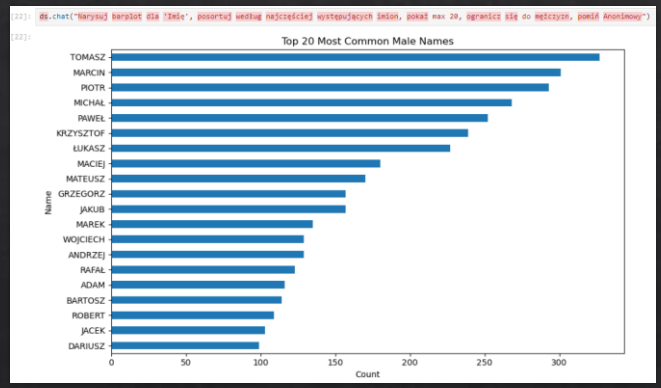
Indywidualna eksploracja każdej kolumny: IMIĘ (M)

```
[19]: ds.chat("Narysuj barplot dla 'Imię', pokaż max 20, ogranicz się do mężczyzn")
```



Wnioski i komentarze

- Najczęściej powtarzające się imiona męskie:
  Tomasz, Marcin, Piotr, Michał…
- Poniżej – ten sam wykres po odfiltrowaniu „anonimów”.
  (z niewiadomych dla mnie przyczyn AI postanowiło zamienić osie)

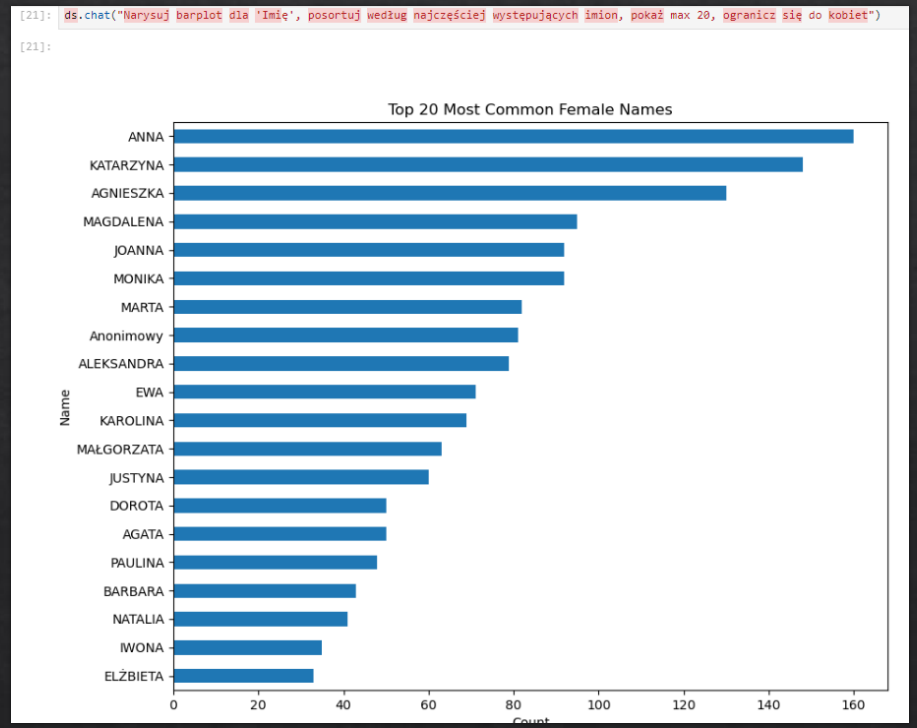


Slajd 4/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: IMIĘ (Ż)

```
[21]: ds.chat("Narysuj barplot dla 'Imię', posortuj według najczęściej występujących imion, pokaż max 20, ogranicz się do kobiet")
```

[21]:

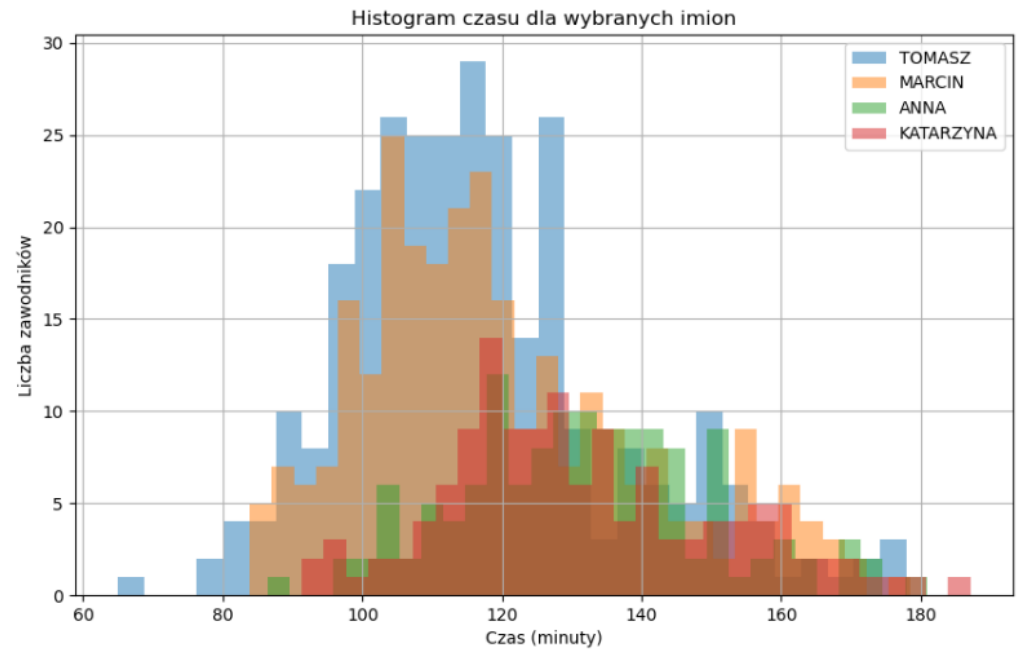


Wnioski i komentarze

- Najczęściej powtarzające się imiona żeńskie: Anna, Katarzyna, Agnieszka, Magdalena…

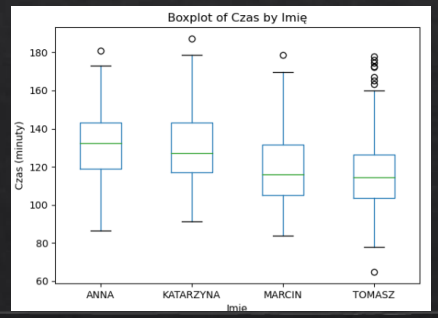
Slajd 5/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: IMIĘ (Ż)



Wnioski i komentarze

- Pojawia się jednak pytanie, kiedy mamy największą szansę, że ktoś nam odpowie na okrzyk „Brawo <imię>?
- Proponuję następujące podejście (chociaż nie jest to jedyna metoda):
  - Wybierzmy po dwa imiona o najliczniejszej grupie kategorii męskiej i żeńskiej . Są to TOMASZ, MARCIN, ANNA i KATARZYNA.
  - Zobaczmy, w jakiej minucie na metę przybiega najliczniejsza grupa biegaczy o tych imionach
  - Załączony histogram i bloxplot pokazują w jakiej minucie maratonu na mecie pojawiła się najliczniejsza grupa Tomków, Marcinów, Ann i Katarzyn.



Uwaga:

Kibicować należało przez cały czas. Tu jedynie pokazuję, kiedy była największa szansa na odzew (że ktoś nam odmacha).

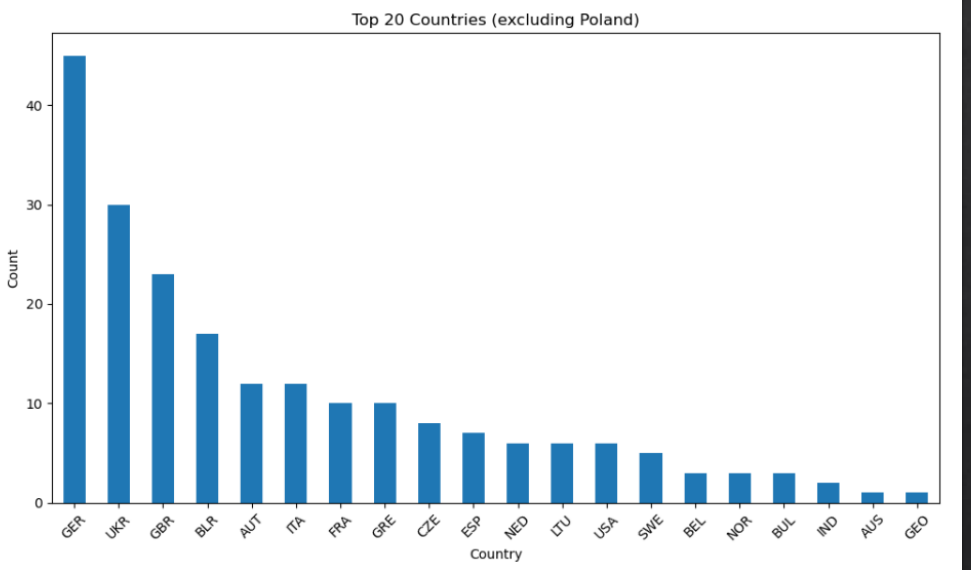
Oczywiście ma to sens, gdy powtórzymy maraton w podobnym składzie i zawodnicy będą
w podobnej formie, chociaż nieco starsi….

Slajd 6/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: KRAJ, z wyłączeniem Polski

```
[23]: ds.chat("Narysuj barplot dla 'Kraj', posortuj według najczęściej występujących krajów, pomiń Polskę")
```

[23]:

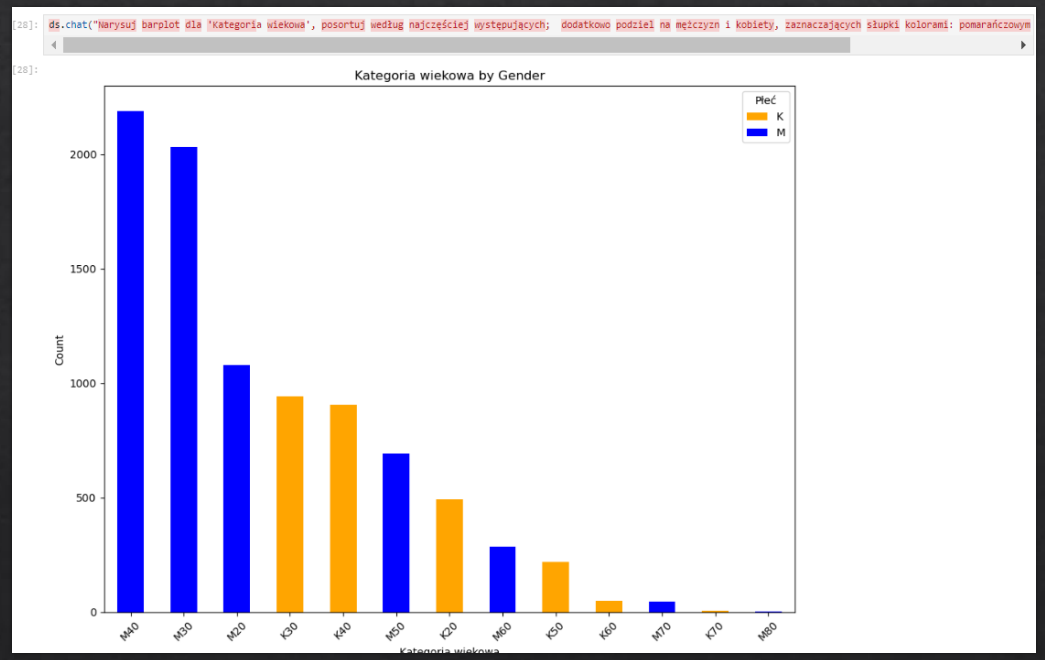


Wnioski i komentarze

- Po odfiltrowaniu zawodników z Polski okazało się, że jest jeszcze grupa uczestników reprezentujących 20 innych krajów.

Slajd 7/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: KATEGORIA WIEKOWA

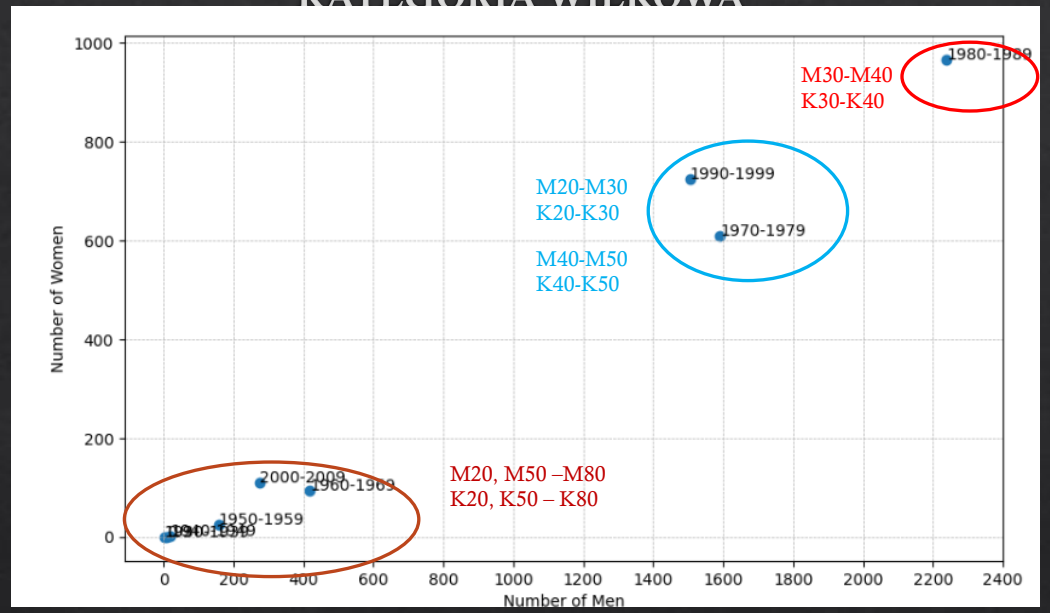


Wnioski i komentarze

- W tym półmaratonie pobiegło zaskakująco dużo mężczyzn z najmłodszej kategorii wiekowej M20. Oczywiście najwięcej z nich było z kategoriiM30 i M40. To samo można powiedzieć o grupie kobiet (K20, K30, K40), chociaż w każdym przedziale było ich o połowę mniej (w przybliżeniu).
- Proporcje te są zakłócone dopiero w starszych kategoriach M,K60, M,K70, M,K80 (!) na korzyść zwiększonej ilości mężczyzn.

Slajd 8/9 ~ Krok 3: Analiza zmiennych

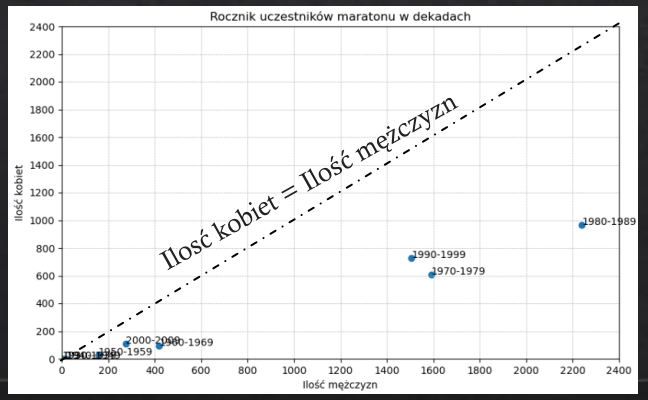
Indywidualna eksploracja każdej kolumny: KATEGORIA WIEKOWA



- Trend: (1) wzrastające zaintersowanie średniego pokolenia, (2) potem „pik” dziewiątej dekady XXw, (3) mniejsze zaangażowanie dekady dziesiątej XXw., (4) małe zaintersowanie pierwszej dekady XXI w.

Wnioski i komentarze

- Sytuację tę ilustruje wykres scatter plot (udało się narysować jednym promptem!). Jednak tutaj podział na grupy wiekowe został zmieniony, ponieważ AI narysowała wykres według dekad (o co zresztą prosiłem).
- Dało to jednak nieoczekiwany efekt. Z wykresu widać, że najliczniejszą grupą byli biegacze urodzeni w dziewiątej dekadzie XX wieku. Ludzie urodzeni po 2000 stanowili dość małą grupę, typową co do liczności bardziej dla seniorów (M50, K50). Średnią co do liczności grupę reprezentowali przedstawiciele pokolenia ostatniej dekady XX wieku i młodsi pięćdziesiątlatkowie ze starszymi czterdziestolatkami. Wyrównanie osi pomaga uchwycić proporcje wg. płci (poniżej)



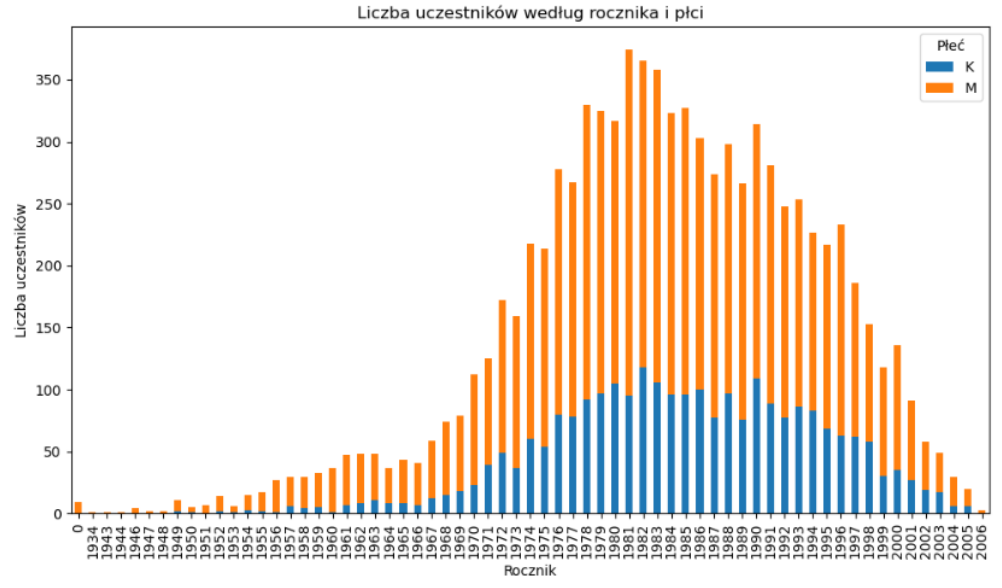
Slajd 9/9 ~ Krok 3: Analiza zmiennych

Indywidualna eksploracja każdej kolumny: KATEGORIA WIEKOWA

```
ds.chat("Narysuj barplot dla 'Rocznik', posortuj według rocznika, dodatkowo podziel na mężczyzn i kobiety")
```

```
<string>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```



Wnioski i komentarze

```
ds.chat("Jaki jest najstarszy rocznik biegacza? Pomin 0.0")
```

```
1934.0
```

```
[41]: ds.chat("Które miejsce zajął biegacz znajstarszym rocznikiem? Pomin 0.0")
```

```
[41]: 8137.0
```

Prawdopodnie Chat GPT ma trudności ze zrozumieniem osi czasu.

```
[42]: ds.chat("Jaki jest najmłodszy rocznik biegacza i które miejsce zajął?")
```

```
[42]: "The youngest runner's birth year is 1934.0 and they placed 8137.0."
```

Teraz lepiej…. Warto uważać na sposób sformułowania prompt'a.

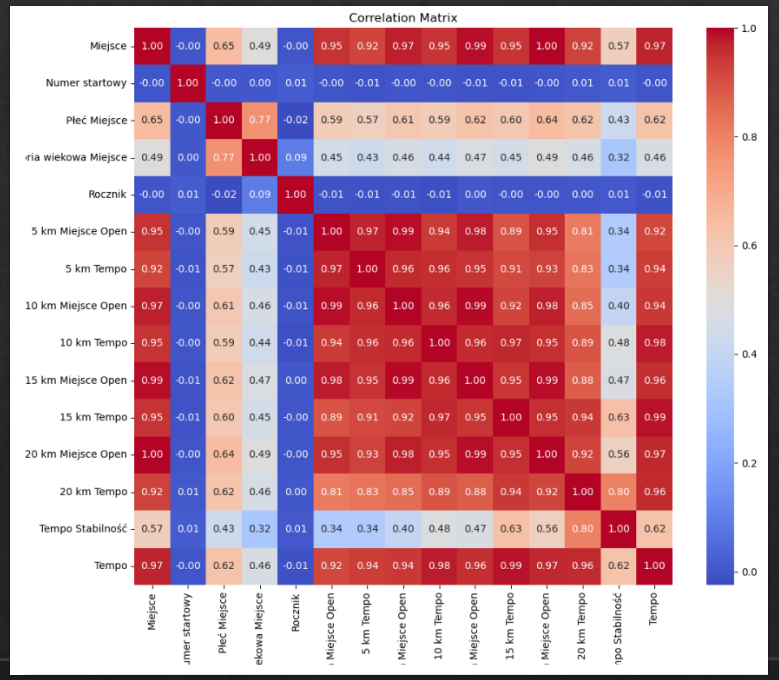
```
[44]: ds.chat("Jaka jest największa wartość rocznika i które miejsce zajął jego posiadacz?")
```

```
[44]: 'The highest birth year is 2006.0 and the place is 1604.0.'
```

Slajd 0/0 ~ Krok 4 – pominięty dla tego przypadku
Slajd 1/1 ~ Krok 5 – Analiza relacji pomiędzy zmiennymi

Macierz korelacji dla kolumn

```
ds.chat("Narysuj macierz korelacji dla kolumn")
```



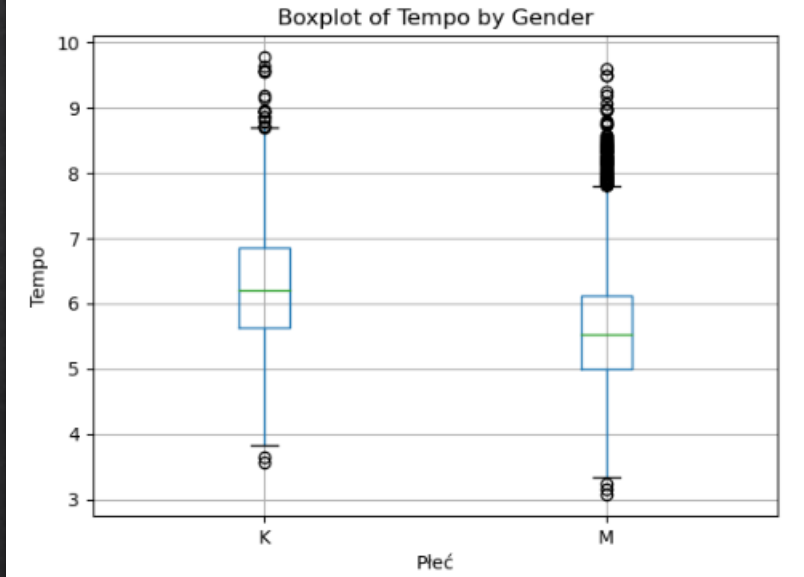
Wnioski i komentarze

- Widoczna jest duża korelacja „Miejsca", z „Tempem" i czasami odcinkowymi oraz powiązanymi z nimi Miejscami typu „Open". Parametry te są ze sobą wielokierunkowo skorelowane. Jest to sytuacja standardowa.
- Nie ma korelacji pomiędzy „Miejscem", a „Numerem Startowym". To również można uznać za sytuację standardową. Dopiero w następnym roku (2024) lepsi biegacze prawdopodobnie zaczęli się rejestrować wcześniej, uzyskując tym sposobem niższe numery. Stąd wystąpiła „sezonowa" korelacja (jednak nie wbudowana na trwale w strukturę procesu).
- Warto zauważyć, że „Tempo Stabilność" jest skorelowane z „Tempem" – co prawda słabiej, ale dostatecznie silnie, by wnioskować :„większa stabilność daje w rezultacie większe tempo" (bo chyba nie na odwrót).

Slajd 1/2 ~ Krok 6 – Poszukiwanie wartości odstających

Bloxplot dla mężczyzn i dla kobiet

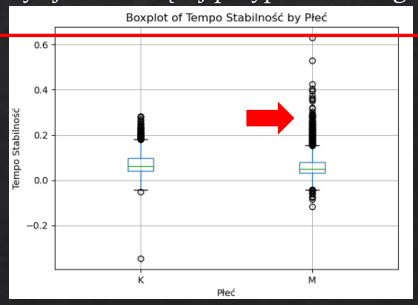
```
ds.chat("Narysuj boxplot dla 'Tempo', rozdziel ze względu na Płeć")
```



Wnioski i komentarze

- Widać trzy grupy uczestników:
  - „zawodowcy" – wytrenowani „robią wynik" nieosiągalny dla reszty. Są wśród nich mężczyźni i kobiety.
  - Potem biegnie główna część wyścigu miarę zwartym tempem (przy czym u mężczyzn jest ono nieco większe).
  - Na końcu biegną prawdziwi bohaterowie wyścigu, pokonujący swoje ograniczenia. Biegną własnym tempem i dobiegają do mety.
  - Bloxplot „Tempo stabilność pokazuje, że zarówno mężczyźni, jak i kobiety pokonali półmaraton w większości stabilnym tempem.

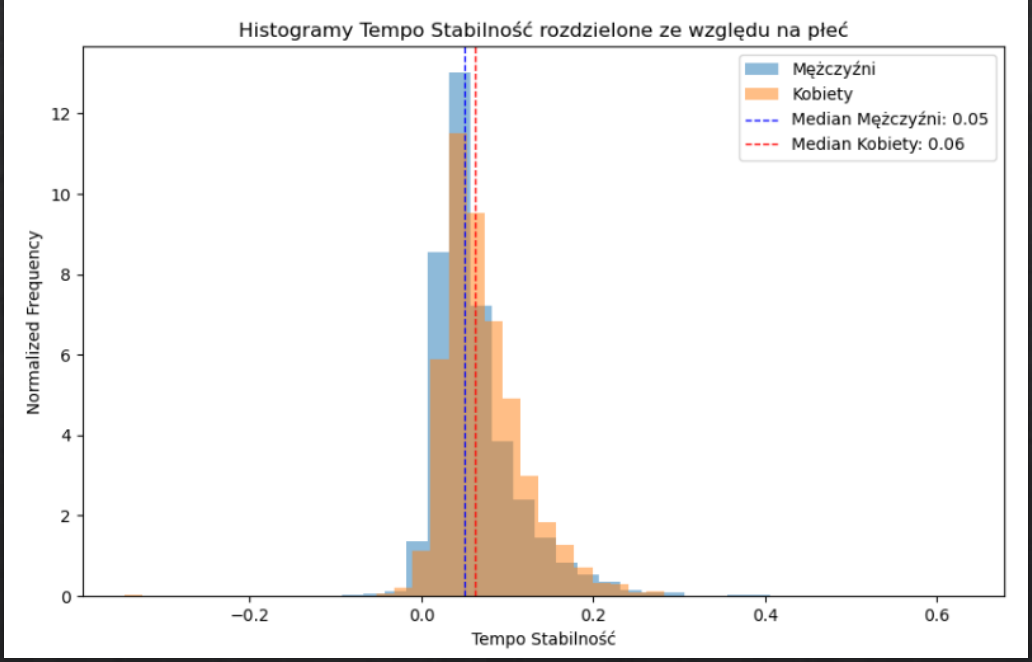
Czy w grupie mężczyzn mogło być jednak więcej przypadków biegaczy spowalniających tempo?



Lepiej jest zadać pytanie, niż pochopnie wyciągać wnioski

Slajd 2/2 ~ Krok 6 – Poszukiwanie wartości odstających

TEMPO STABILNOŚĆ rozdzielone dla mężczyzn i dla kobiet



Wnioski i komentarze

- Sytuację dobrzej wyjaśnia histogram.
- Grupa „outliers” (wartości odstających) jest bardziej zwarta w przypadku kobiet. Dlatego na wykresie typu bloxplot sprawiała wrażenie mniej licznej.