

Analiza wyników w bieganiu na dystansach olimpijskich, półmaratonie oraz maratonie

Bartłomiej Machura

PRZETWARZANIE ANALIZA, WIZUALIZACJA
DANYCH W SYSTEMIE SAS



Uniwersytet
Ekonomiczny
w Katowicach

Spis treści

| | |
|---|----|
| Wprowadzenie do problematyki | 3 |
| Opis źródeł danych | 4 |
| Analiza osiągniętych wyników ze względu na narodowość zawodnika | 7 |
| Analiza osiągniętych wyników ze względu na wiek zawodnika | 12 |
| Analiza osiągniętych wyników w czasie | 17 |
| Analiza osiągniętych wyników ze względu na płeć zawodnika | 21 |
| Analiza wpływu daty i miejsca zawodów na uzyskiwane wyniki. | 24 |
| Wnioski | 27 |

Wprowadzenie do problematyki

W dzisiejszym dynamicznym społeczeństwie, coraz większą uwagę przykładamy do zdrowego trybu życia, aktywności fizycznej i dbania o kondycję fizyczną. Jedną z najbardziej popularnych form aktywności jest bieganie, które nie tylko dostarcza wielu korzyści dla zdrowia fizycznego, ale również przyczynia się do poprawy kondycji psychicznej.

Cele szczegółowe:

- **Analiza osiągniętych wyników ze względu na narodowość zawodnika**
- **Analiza osiągniętych wyników ze względu na wiek zawodnika**
- **Analiza osiągniętych wyników w czasie (historycznie)**
- **Analiza osiągniętych wyników ze względu na płeć zawodnika**
- **Analiza wpływu daty i miejsca zawodów na uzyskiwane wyniki.**

W mojej pracy przeanalizuję poszczególne czynniki i ich wpływ na dyspozycje zawodników i uzyskiwane przez nich czasy na poszczególnych dystansach.

Opis źródeł danych

Dane do analizy uzyskałem z serwisu Kaggle.com. Dotyczą one wyników biegowych uzyskiwanych przez najlepszych zawodników na świecie. Zawierają dane o dystansach olimpijskich oraz półmaratonie. Reprezentują one 1000 najlepszych wyników w historii uzyskanych na tych dystansach. Dzięki nim znamy podstawowe dane o zawodniku: Imię, Nazwisko, Kraj pochodzenia, Miejsce ogółem jak i na zawodach, jak i o biegach: Data, miejsce zawodów.

Mój projekt rozpocząłem od utworzenia biblioteki, w której będę umieszczał utworzone przeze mnie tabele. Użyłem do tego poniższego kodu:

```
Libname Projekt '/home/u63619452/Student/PROJEKT';  
run;
```

Kod 1: Utworzenie biblioteki

Dane, które uzyskałem były w formacie CSV, więc następnie zająłem się zaimportowaniem pliku. Wykorzystałem do tego kod poniżej:

```
FILENAME REFFILE '/home/u63619452/Student/csv/data.csv';  
  
PROC IMPORT DATAFILE=REFFILE  
    DBMS=CSV  
    OUT=Projekt.Dane;  
    GETNAMES=YES;  
RUN;
```

KOD 2: Import pliku CSV

| | Rank | Time | Name | Country | Date of Birth | Place | City | Date | Gender | Event |
|----|------|------------|-----------------|---------|---------------|-------|----------------|------------|--------|-------|
| 1 | 1 | 0:01:40.91 | David Rudisha | KEN | 1988-12-17 | 1 | London | 2012-09-08 | Men | 800 m |
| 2 | 2 | 0:01:41.01 | David Rudisha | KEN | 1988-12-17 | 1 | Rieti | 2010-08-29 | Men | 800 m |
| 3 | 3 | 0:01:41.09 | David Rudisha | KEN | 1988-12-17 | 1 | Berlin | 2010-08-22 | Men | 800 m |
| 4 | 4 | 0:01:41.11 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Köln | 1997-08-24 | Men | 800 m |
| 5 | 5 | 0:01:41.24 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Zürich | 1997-08-13 | Men | 800 m |
| 6 | 6 | 0:01:41.33 | David Rudisha | KEN | 1988-12-17 | 1 | Rieti | 2011-10-09 | Men | 800 m |
| 7 | 7 | 0:01:41.51 | David Rudisha | KEN | 1988-12-17 | 1 | Heusden-Zolder | 2010-10-07 | Men | 800 m |
| 8 | 8 | 0:01:41.54 | David Rudisha | KEN | 1988-12-17 | 1 | Saint-Denis | 2012-06-07 | Men | 800 m |
| 9 | 9 | 0:01:41.73 | Sebastian Coe | GBR | 1956-09-29 | 1 | Firenze | 1981-10-06 | Men | 800 m |
| 10 | 9 | 0:01:41.73 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Stockholm | 1997-07-07 | Men | 800 m |

Zaimportowane dane

Następnie pozmieniałem nazwy kolumn na polskie za pomocą kodu nr 3.

```

Data Projekt.DaneZ;
set Projekt.Dane;
Rename Rank = Ranking;
Rename Time=Czas;
Rename Name= Zawodnik;
Rename 'Date of Birth' = Data_urodzenia;
Rename Place=Miejsce;
Rename Country=Kraj;
Rename City=Miasto;
Rename Date=Data;
Rename Gender=Płeć;
Rename Event=Dystans;
run;

```

Kod 3: Zmiana nazw kolumn

| | Ranking | Czas | Zawodnik | Kraj | Data_urodzenia | Miejsce | Miasto | Data | Płeć | Dystans |
|----|---------|------------|-----------------|------|----------------|---------|----------------|------------|------|---------|
| 1 | 1 | 0:01:40.91 | David Rudisha | KEN | 1988-12-17 | 1 | London | 2012-09-08 | Men | 800 m |
| 2 | 2 | 0:01:41.01 | David Rudisha | KEN | 1988-12-17 | 1 | Rieti | 2010-08-29 | Men | 800 m |
| 3 | 3 | 0:01:41.09 | David Rudisha | KEN | 1988-12-17 | 1 | Berlin | 2010-08-22 | Men | 800 m |
| 4 | 4 | 0:01:41.11 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Köln | 1997-08-24 | Men | 800 m |
| 5 | 5 | 0:01:41.24 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Zürich | 1997-08-13 | Men | 800 m |
| 6 | 6 | 0:01:41.33 | David Rudisha | KEN | 1988-12-17 | 1 | Rieti | 2011-10-09 | Men | 800 m |
| 7 | 7 | 0:01:41.51 | David Rudisha | KEN | 1988-12-17 | 1 | Heusden-Zolder | 2010-10-07 | Men | 800 m |
| 8 | 8 | 0:01:41.54 | David Rudisha | KEN | 1988-12-17 | 1 | Saint-Denis | 2012-06-07 | Men | 800 m |
| 9 | 9 | 0:01:41.73 | Sebastian Coe | GBR | 1956-09-29 | 1 | Firenze | 1981-10-06 | Men | 800 m |
| 10 | 9 | 0:01:41.73 | Wilson Kipketer | DEN | 1970-12-12 | 1 | Stockholm | 1997-07-07 | Men | 800 m |

Dane po zmienieniu nazw kolumn

Następnie przy użyciu kodu nr 4 utworzyłem kolumnę zawierającą dystans w metrach z typem numerycznym.

```

Data Projekt.DaneG;
set Projekt.DaneZ;

If Dystans ="10,00" then DystansM = 10000;
If Dystans ="800 m" then DystansM = 800;
If Dystans ="400 m" then DystansM = 400;
If Dystans ="Marat" then DystansM = 42195;
If Dystans ="100 m" then DystansM= 100;
If Dystans ="Half" then DystansM= 21097;
If Dystans ="200 m" then DystansM = 200;
If Dystans ="1500" then DystansM = 1500;
If Dystans ="5000" then DystansM = 5000;
Keep Ranking Czas Zawodnik Kraj Data_urodzenia Miejsce Miasto Data Płeć DystansM
run;

```

Kod 4: Utworzenie kolumny DystansM

Aby sprawdzić poprawność i logiczność analizowanych danych postanowiłem sprawdzić, czy istnieje oczywista zależność między czasem, dystansem. W tym celu użyłem kodu nr 5.

```
proc corr data = Projekt.Daneg;
VAR DystansM Czas;
run;
```

Kod nr 5: Korelacja Pearsona

Procedura CORR

2 Zmienne: DystansM Czas

| Statystyki proste | | | | | | |
|-------------------|-------|---------|------------|-----------|-----------|----------|
| Zmienna | N | Średnia | Odch. std. | Suma | Minimum | Maksimum |
| DystansM | 18244 | 8922 | 13356 | 162767368 | 100.00000 | 42195 |
| Czas | 18244 | 1648 | 2579 | 30073850 | 9.58000 | 8784 |

Współczynniki korelacji Pearsona, N = 18244
Prawd. > |r| przy H0: rho=0

| | DystansM | Czas |
|----------|--------------------|--------------------|
| DystansM | 1.00000 | 0.99667 < .0001 |
| Czas | 0.99667 < .0001 | 1.00000 |

Jak się można było spodziewać wartość korelacji jest bliska 1, co potwierdza, iż dane są właściwe.

Analiza osiągniętych wyników ze względu na narodowość zawodnika

Celem tej analizy jest pokazanie rozkładu ilości zawodników na poszczególnych dystansach ze względu na narodowość. Taka analiza to złożony proces, który wymaga uwzględnienia wielu czynników. Każdy sportowiec jest jednostką unikatową, jednak analiza narodowości może pomóc w zrozumieniu szerszego kontekstu, w jakim rozwijają się i osiągają sukcesy. Czynnikiem wpływającym na wyniki biegaczy mogą być m.in. takie elementy jak: historia sportowa i infrastruktura sportowa w danym kraju, wsparcie finansowe dla sportowców, system szkoleniowy, kultura sportowa, genetyczne predyspozycje i cechy fizyczne oraz polityka i struktury sportowe w danym kraju.

Postanowiłem to sprawdzić używając kodu nr 6.

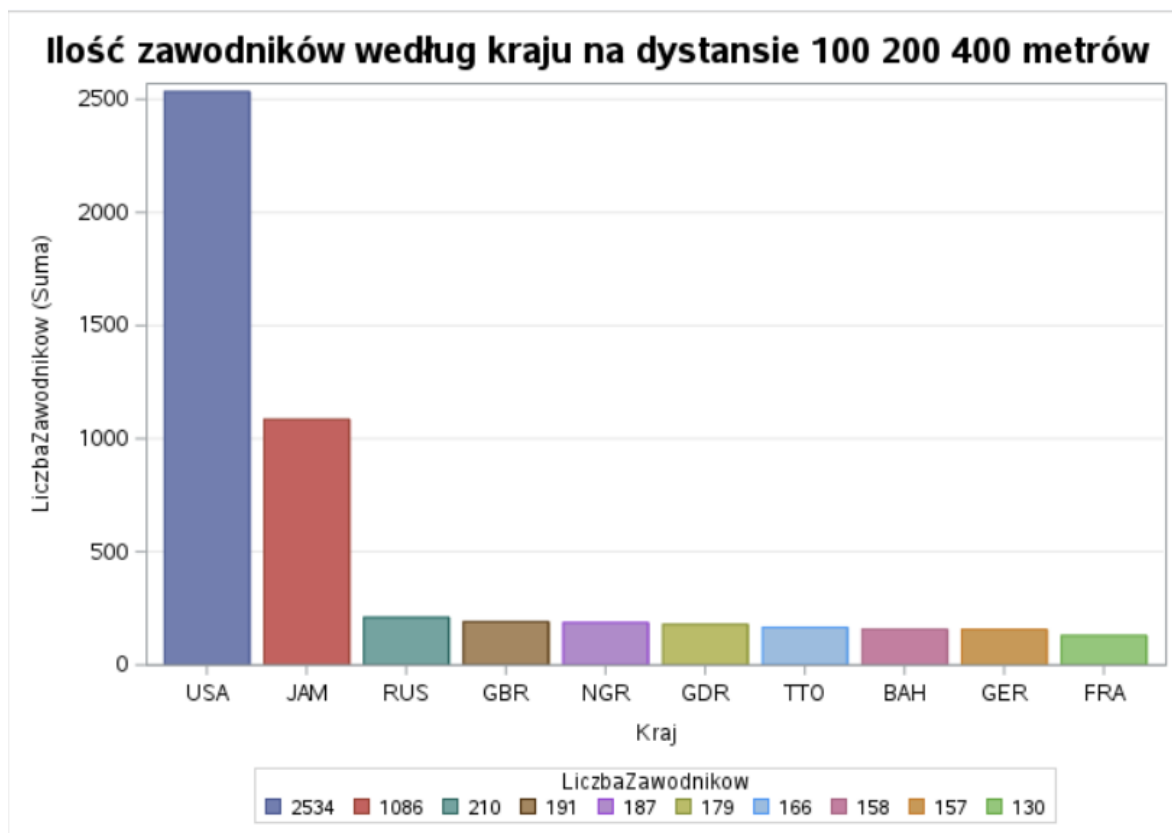
```
proc sort data = Projekt.daneg;
by kraj;
run;
data Analiza1;
    set Projekt.daneg;
    where DystansM = 100 or DystansM = 200 or DystansM = 400;
    by Kraj;
    if first.Kraj then LiczbaZawodnikow = 1;
    else LiczbaZawodnikow + 1;
    if last.Kraj then output;
    Keep Kraj LiczbaZawodnikow;
run;
proc sort data = Analiza1;
by LiczbaZawodnikow;
run;
proc gchart data=Analiza1;
    title1 'Ilość zawodników według kraju na dystansie 100 200 400 metrów';
    vbar Kraj / sumvar=LiczbaZawodnikow discrete ;
run;
```

Kod 6: Pogrupowanie danych i stworzenie wykresu.

Postanowiłem pogrupować dystans na kilka kategorii:

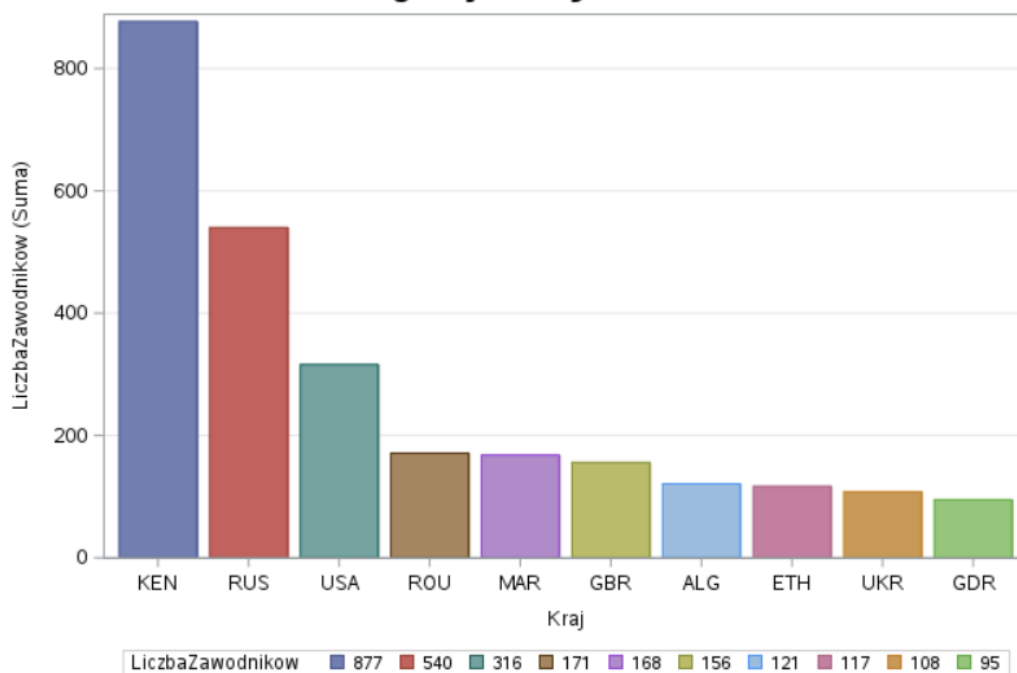
- Biegi krótkie – 100 200 oraz 400 m
- Biegi średnie 800 oraz 1500 m
- Biegi długie 5 oraz 10 km
- Półmaraton oraz maraton

Następnie utworzyłem wykresy słupkowe pokazujące rozkład ilości zawodników na poszczególnych dystansach.



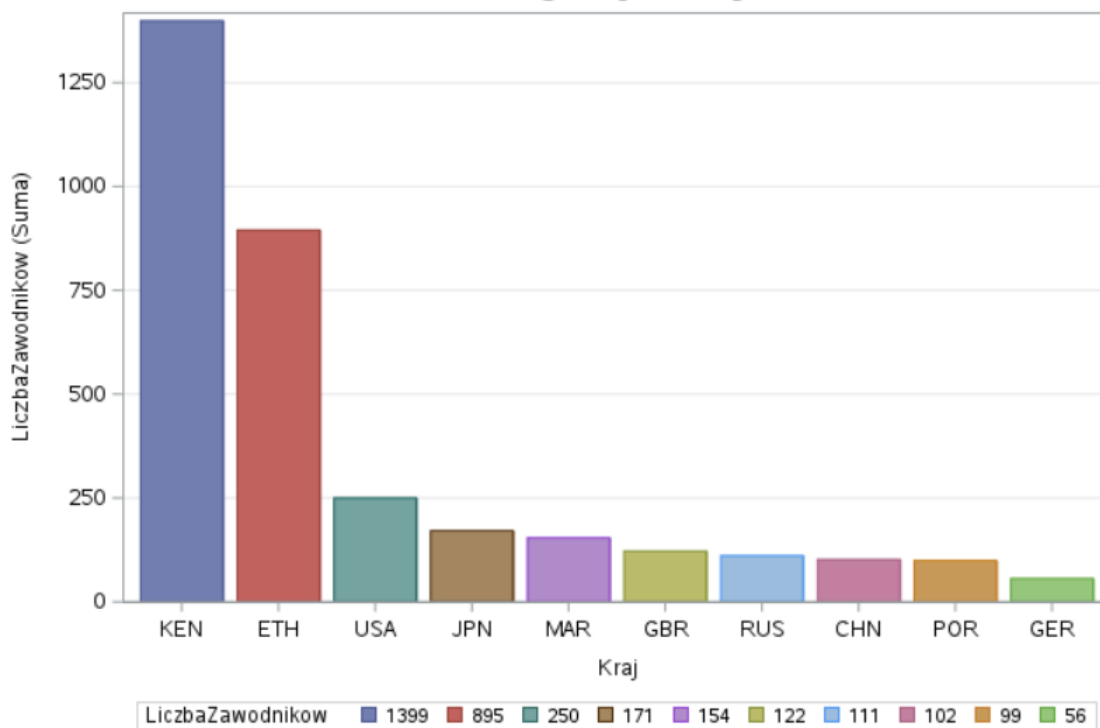
Jak widać na dystansach sprinterskich 100, 200 i 400 metrów zdecydowaną przewagę ma USA. Następny kraj to Jamajka, skąd pochodzi najlepszy sprinter w historii Usain Bolt. Trzecie miejsce na tym dystansie zajmuje Rosja.

Ilość zawodników według kraju na dystansie 100 200 400 metrów

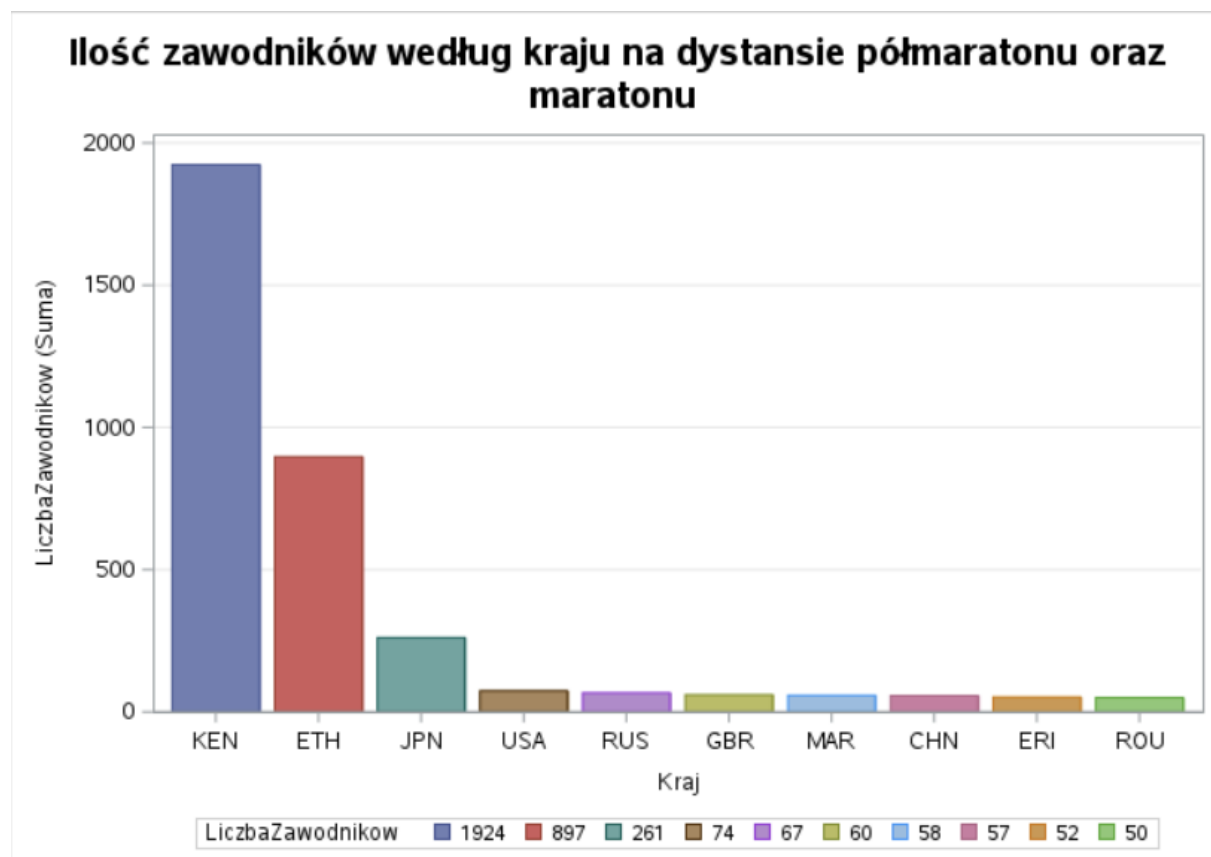


W przypadku 800 oraz 1500 metrów widać przewagę Kenijczyków. Zaraz za nimi znajdują się zawodnicy z Rosji oraz Stanów Zjednoczonych.

Ilość zawodników według kraju na dystansie 5 oraz 10 km



Na dystansach 5 km oraz 10 km utrzymuje się pozycja Kenii oraz pojawia się następne państwo Etiopia. 3 Miejsce ponownie zajmują Stany Zjednoczone.



Na dystansie maratonu ponownie dominują Kenijczycy oraz Etiopczycy. 3 miejsce zajmuje tu natomiast Japonia.

Patrząc na te wykresy można dojść do wniosku, że są kraje, które posiadają zawodników osiągających lepsze wyniki. Jak pokazała analiza dystans też jest przy tym kluczowy. I tak w przypadku sprinterów najlepsze wyniki osiągają osoby z USA, Jamajki oraz Rosji. Może to być związane z dość dużą skutecznością dopingu na tym dystansie. Wszystkie te kraje mają za sobą niejedną taką aferę. W przypadku średnich dystansów do grona krajów z najlepszymi zawodnikami dołącza Kenia, która jako lider utrzyma się już na wszystkich dystansach. Na dystansach 5 i 10 km spada udział USA i Rosji, a na 2 miejscu pojawia się Etiopia. Pokazuje to, iż wraz z wydłużaniem się dystansu przewagę zyskują zawodnicy czarnoskórzy. Wpływ na taki stan rzeczy ma kilka czynników min. genetyka, warunki treningowe, system szkolenia, motywacja psychologiczna. Wiele osób zaczyna w tych krajach trening już nawet w wieku 6 lat, a sukces jest dla nich jedyną możliwością by wyjść z biedy. Tak więc na dystansie półmaratonu oraz maratonu nadal liderem pozostaje Kenia, tuż za nią znajdują się Etiopczycy, a co ciekawe na 3 miejscu pojawia się Japonia. Może to być

związane z dużą aktywnością fizyczną Japończyków, szczególnie jeśli chodzi o bieganie, jak i wieloma zawodami dla uczniów szkół wyższych.

Analiza osiągniętych wyników ze względu na wiek zawodnika

Analiza osiągniętych wyników biegowych ze względu na wiek zawodnika jest kluczowym narzędziem w zrozumieniu dynamiki rozwoju sportowca oraz jego zdolności adaptacyjnych na różnych etapach życia. Badanie wyników w kontekście wieku może wykazać, jak doświadczenie wpływa na osiągnięte rezultaty. Czy sportowiec osiąga szczyt swojej kariery w konkretnym wieku, czy też doświadczenie przekłada się na długotrwałą konsystencję wyników. Pozwala również śledzić ewolucję umiejętności zawodnika. Czy jego osiągnięcia poprawiają się, czy też istnieją okresy, w których można zauważyć wyraźny skok jakościowy. Aby odpowiedzieć na te pytania sprawdzę w jakim wieku byli zawodnicy, kiedy osiągnęli najlepsze swoje wyniki. W tym celu utworzyłem zmienną Wiek, zawierającą wiek zawodnika w dniu zawodów.

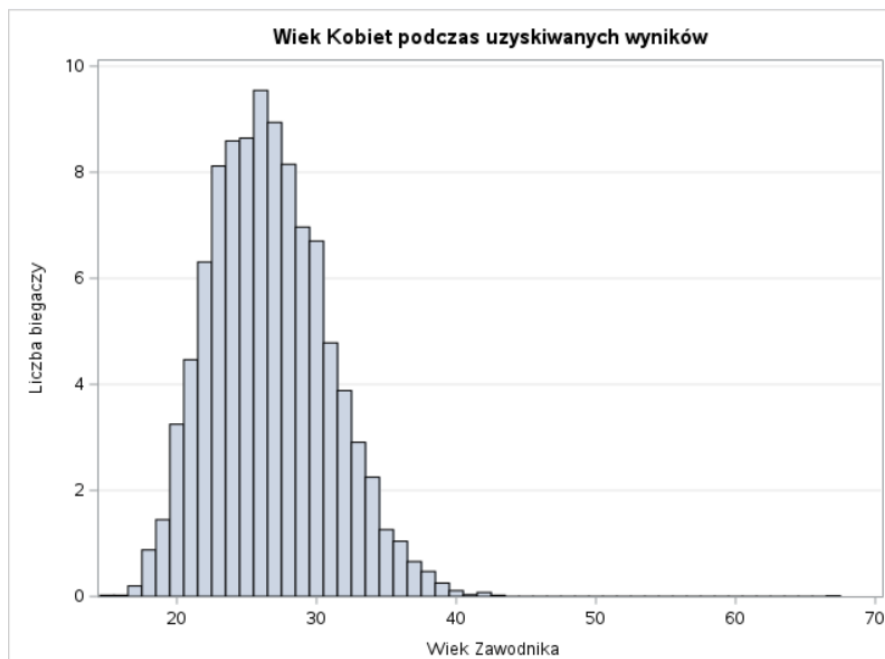
```
Data Projekt.Analiza2;  
set Projekt.daneg;  
Wiek = intck('year', Data_urodzenia, Data);  
run;
```

Kod nr 7: Utworzenie zmiennej Wiek.

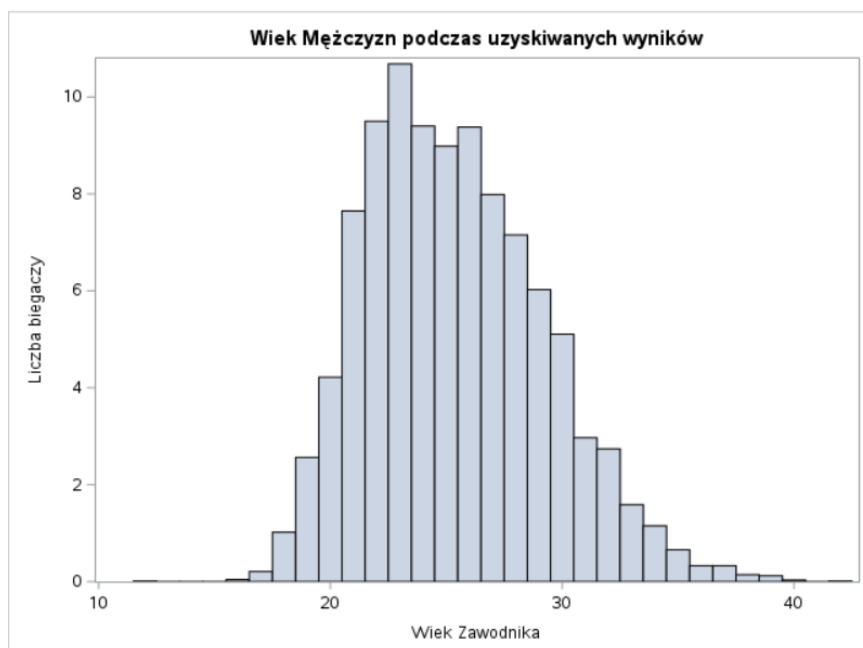
Postanowiłem, że sprawdzę z podziałem na płcie w jakim wieku zawodnicy uzyskiwali swoje wyniki. Użyłem do tego histogramu stworzonego za pomocą kodu nr 8.

```
proc sgplot data=Projekt.Analiza2;  
  histogram Wiek / binwidth=1;  
  title1 'Wiek Kobiety podczas uzyskiwanych wyników';  
  xaxis label="Wiek Zawodnika";  
  yaxis label="Liczba biegaczy" grid;  
run;
```

Kod nr 8: Utworzenie histogramu



Histogram pokazuje, że pomiędzy 20 a 35 rokiem kobiety uzyskują swoje najlepsze wyniki. Statystycznie najlepszy wiek, aby pobijać rekordy to 26 lat.



W przypadku mężczyzn 23 lata to wiek, w którym osiągają najlepsze wyniki. W wieku 20 lat widać wzrost jakości ich biegania, a przy zbliżaniu się do 30 roku życia spadek osiąganych przez nich wyników.

| Wiek Kobiet podczas uzyskiwanych wyników | | | | | Wiek Mężczyzn podczas uzyskiwanych wyników | | | | |
|--|------------|---------|------------------------|---------------------|--|------------|---------|------------------------|---------------------|
| Procedura FREQ | | | | | Procedura FREQ | | | | |
| Wiek | Liczebność | Procent | Liczebność skumulowana | Procent skumulowany | Wiek | Liczebność | Procent | Liczebność skumulowana | Procent skumulowany |
| 15 | 2 | 0.02 | 2 | 0.02 | 12 | 1 | 0.01 | 1 | 0.01 |
| 16 | 2 | 0.02 | 4 | 0.04 | 16 | 4 | 0.04 | 5 | 0.05 |
| 17 | 18 | 0.20 | 22 | 0.24 | 17 | 19 | 0.21 | 24 | 0.26 |
| 18 | 80 | 0.88 | 102 | 1.12 | 18 | 93 | 1.02 | 117 | 1.28 |
| 19 | 132 | 1.45 | 234 | 2.57 | 19 | 234 | 2.56 | 351 | 3.85 |
| 20 | 296 | 3.25 | 530 | 5.81 | 20 | 385 | 4.22 | 736 | 8.06 |
| 21 | 407 | 4.46 | 937 | 10.28 | 21 | 698 | 7.65 | 1434 | 15.71 |
| 22 | 575 | 6.31 | 1512 | 16.59 | 22 | 867 | 9.50 | 2301 | 25.21 |
| 23 | 740 | 8.12 | 2252 | 24.70 | 23 | 975 | 10.68 | 3276 | 35.89 |
| 24 | 783 | 8.59 | 3035 | 33.29 | 24 | 858 | 9.40 | 4134 | 45.29 |
| 25 | 788 | 8.64 | 3823 | 41.94 | 25 | 820 | 8.98 | 4954 | 54.28 |
| 26 | 870 | 9.54 | 4693 | 51.48 | 26 | 856 | 9.38 | 5810 | 63.66 |
| 27 | 815 | 8.94 | 5508 | 60.42 | 27 | 729 | 7.99 | 6539 | 71.64 |
| 28 | 743 | 8.15 | 6251 | 68.57 | 28 | 653 | 7.15 | 7192 | 78.80 |
| 29 | 635 | 6.97 | 6886 | 75.54 | 29 | 550 | 6.03 | 7742 | 84.83 |
| 30 | 611 | 6.70 | 7497 | 82.24 | 30 | 466 | 5.11 | 8208 | 89.93 |
| 31 | 436 | 4.78 | 7933 | 87.02 | 31 | 271 | 2.97 | 8479 | 92.90 |
| 32 | 354 | 3.88 | 8287 | 90.91 | 32 | 250 | 2.74 | 8729 | 95.64 |
| 33 | 265 | 2.91 | 8552 | 93.81 | 33 | 145 | 1.59 | 8874 | 97.23 |
| 34 | 205 | 2.25 | 8757 | 96.06 | 34 | 105 | 1.15 | 8979 | 98.38 |
| 35 | 115 | 1.26 | 8872 | 97.32 | 35 | 60 | 0.66 | 9039 | 99.04 |
| 36 | 95 | 1.04 | 8967 | 98.37 | 36 | 30 | 0.33 | 9069 | 99.36 |
| 37 | 60 | 0.66 | 9027 | 99.02 | 37 | 30 | 0.33 | 9099 | 99.69 |
| 38 | 43 | 0.47 | 9070 | 99.50 | 38 | 13 | 0.14 | 9112 | 99.84 |
| 39 | 23 | 0.25 | 9093 | 99.75 | 39 | 11 | 0.12 | 9123 | 99.96 |
| 40 | 10 | 0.11 | 9103 | 99.86 | 40 | 3 | 0.03 | 9126 | 99.99 |
| 41 | 3 | 0.03 | 9106 | 99.89 | 42 | 1 | 0.01 | 9127 | 100.00 |
| 42 | 7 | 0.08 | 9113 | 99.97 | | | | | |
| 43 | 2 | 0.02 | 9115 | 99.99 | | | | | |
| 67 | 1 | 0.01 | 9116 | 100.00 | | | | | |

Analiza ta pokazała, że zdecydowana większość zawodników bez względu na płeć osiąga swoje najlepsze wyniki biegowe w wieku od 20 do 30 lat. Prawdopodobnie jest to związane z wydolnością organizmu, która osiąga swój szczyt między 20-30 rokiem życia, a następnie zaczyna spadać. Statystycznie kobiety osiągają ten szczyt później, co powoduje, że ustanawiają one swoje najlepsze wyniki nieco później od mężczyzn. Postanowiłem, że utworzę przedziały wieku, które obejmą zarówno kobiety jak i mężczyzn i sprawdzę ich procentowy udział. Utworzone do analizy przedziały to:

- do 21 roku życia
- od 22 do 32 roku życia
- powyżej 32 roku życia

```

data Projekt.Analiza2;
set Projekt.Analiza2;
  if Wiek < 21 then Przedział_Wieku = "W1";
  if Wiek > 21 and Wiek <32 then Przedział_Wieku = "W2";
  if Wiek > 32 then Przedział_Wieku = "W3";
run;

```

Kod nr 9: Utworzenie przedziałów wiekowych

Następnie postanowiłem utworzyć wykres kołowy, na którym zaprezentuję procentowy udział wieku osób w podanych przedziałach z podziałem na kobiety i mężczyzn.

```

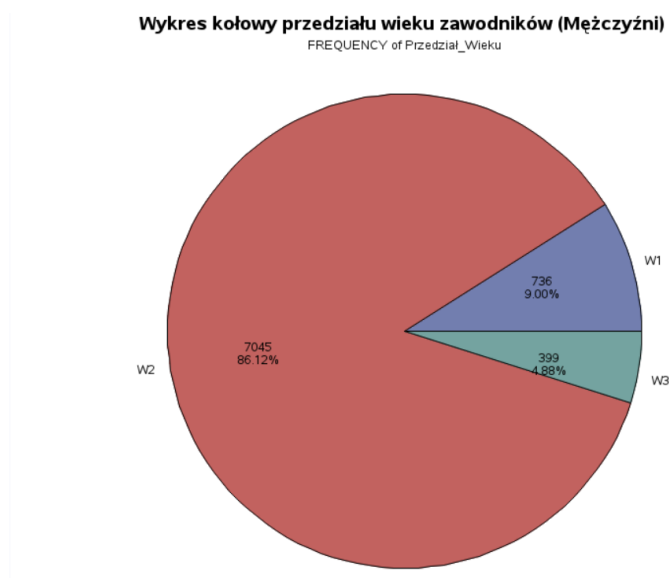
PROC GCHART DATA=Projekt.Analiza2;
PIE Przedział_Wieku / DISCRETE VALUE=INSIDE

PERCENT=INSIDE SLICE=OUTSIDE;

RUN;

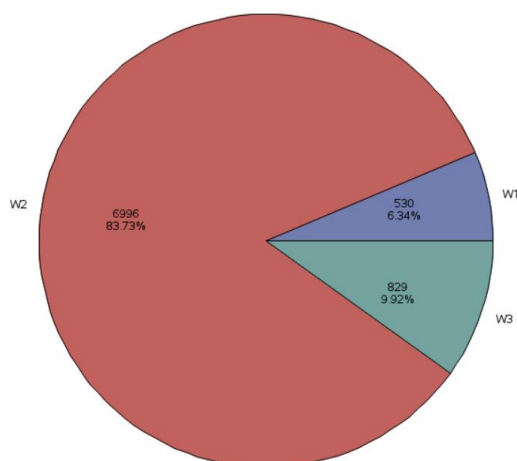
```

Kod nr 10: Utworzenie wykresu kołowego



Wykres kołowy przedziału wieku zawodników (Kobiety)

FREQUENCY of Przedział_Wieku



Procentowy udział wygląda bardzo podobnie, jednak delikatne różnice procentowe potwierdzają tezę, że kobiety później osiągają swoje najlepsze wyniki. Na obu wykresach przedział od 21 do 32 roku życia osiąga ponad 80 procent. Jeśli chodzi o drugi co do wartości procentowej przedział to w przypadku mężczyzn jest to poniżej 21 roku życia, a w przypadku kobiet powyżej 32. Podsumowanie przeprowadzonej analizy dotyczącej osiąganych wyników biegowych ze względu na wiek zawodnika wydaje się dostarczać istotnych informacji na temat dynamiki rozwoju sportowca oraz wpływu doświadczenia na osiągnięte rezultaty.

Wyniki analizy wskazują, że zarówno kobiety, jak i mężczyźni, zdecydowanie częściej osiągają najlepsze wyniki biegowe w przedziale wiekowym od 20 do 30 lat. Statystycznie, najlepszy wiek dla kobiet to około 26 lat, natomiast dla mężczyzn wynosi on 23 lata. To potwierdza, że wydolność organizmu osiąga szczyt między 20 a 30 rokiem życia.

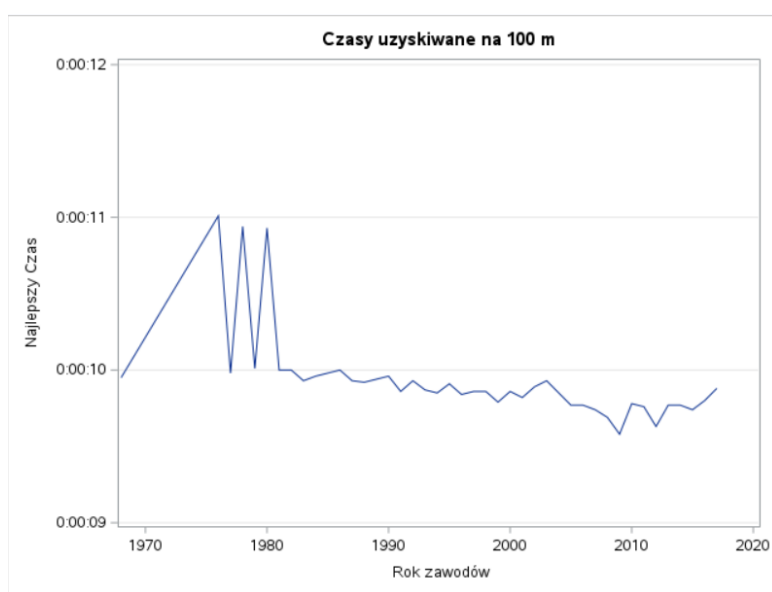
Interesującym spostrzeżeniem jest różnica czasowa między płciami – kobiety zdają się osiągać swój szczyt później niż mężczyźni, co przekłada się na nieco późniejsze ustanawianie najlepszych wyników. Analiza przedziałów wiekowych, obejmujących zarówno kobiety, jak i mężczyzn, potwierdza, że wiek od 21 do 32 roku życia jest kluczowy dla obu płci, stanowiąc ponad 80 procent udziału.

Analiza osiągniętych wyników w czasie

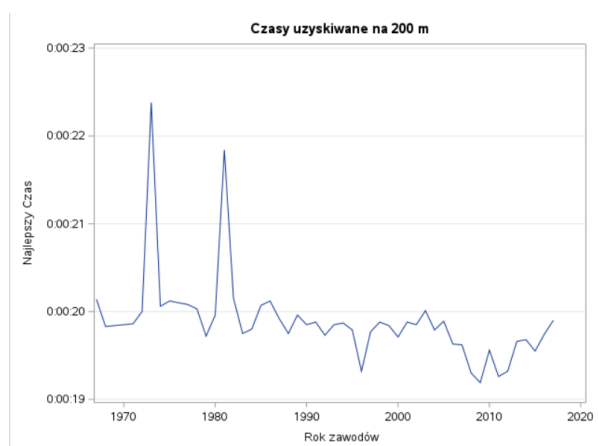
W przypadku tej analizy chciałem sprawdzić, jak przedstawiają się wyniki biegaczy w aspekcie czasowym. Użyłem kodu nr 8, aby sprawdzić, jak wyglądały najlepsze uzyskane wyniki na przestrzeni lat na konkretnych dystansach. Ocena wyników biegowych pozwala śledzić ewolucję techniki biegowej w miarę upływu lat i określić, czy zawodnicy doskonalą swoje umiejętności techniczne lub zmieniają swoje podejście do stylu biegu.

```
Data Projekt.analiza3;  
set Projekt.daneg;  
where DystansM = 100;  
Rok = Year(Data);  
run;  
proc means data=Projekt.analiza3 noprint;  
class Rok;  
var Czas;  
output out=Projekt.czas min=wynik;  
run;  
proc sgplot data=Projekt.czas;  
series x=Rok y=wynik / ;  
title1 "Czasy uzyskiwane na 100 m";  
xaxis label="Rok zawodów" ;  
yaxis label="Najlepszy Czas" grid;  
run;
```

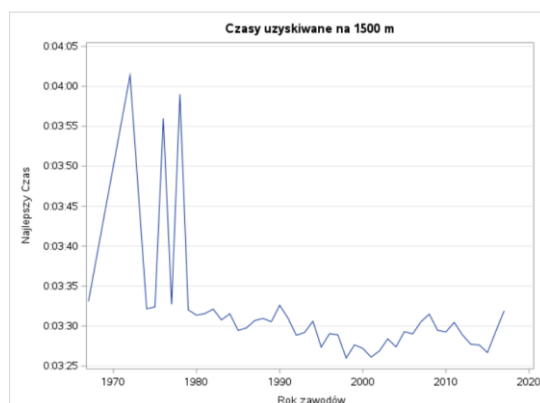
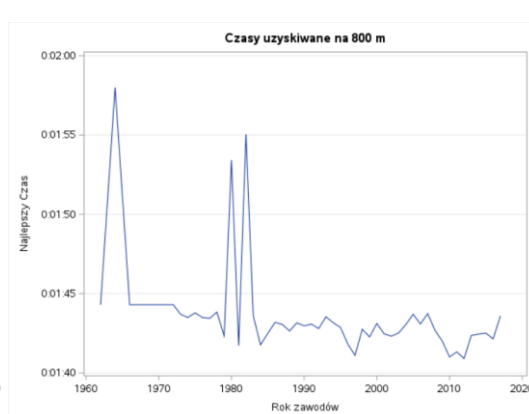
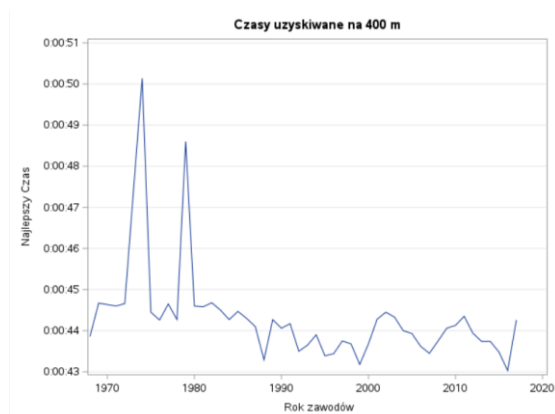
Kod nr 8: Tworzenie wykresu liniowego.



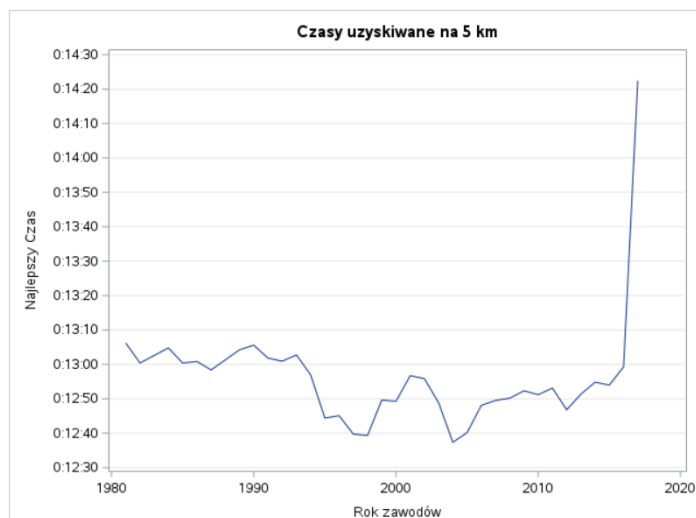
Wykres prezentujący najlepsze wyniki na 100 m pokazuje wahania w latach 1970-1980. Czas wahał się między 10 a 11 sekund. W następnych latach widać stopniową poprawę wyników na przestrzeni lat oraz lekkie pogorszenie wyników w latach 2009-2017. Może być to spowodowane rozwojem kontroli antydopingowej oraz zaostrzeniem jej procedur w tych latach.



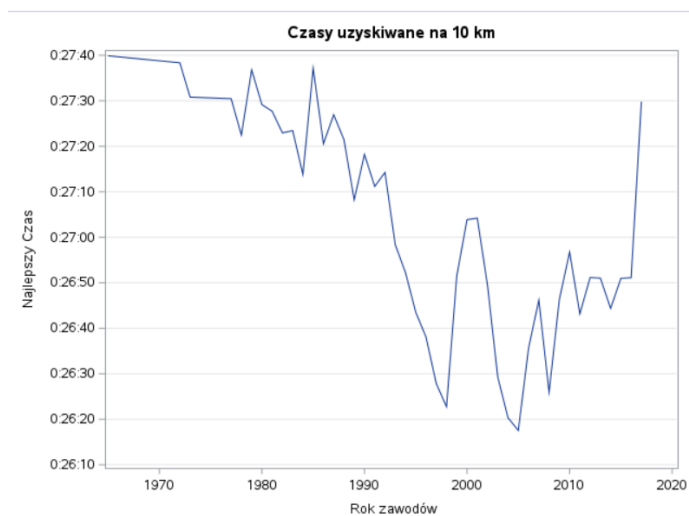
W przypadku 200 metrów sytuacja przypomina dystans 100 metrów. Jednak tutaj wahania w latach 1970-1980 są większe, następnie następuje stopniowa poprawa wyników oraz ich lekki spadek w okolicach 2011 roku.



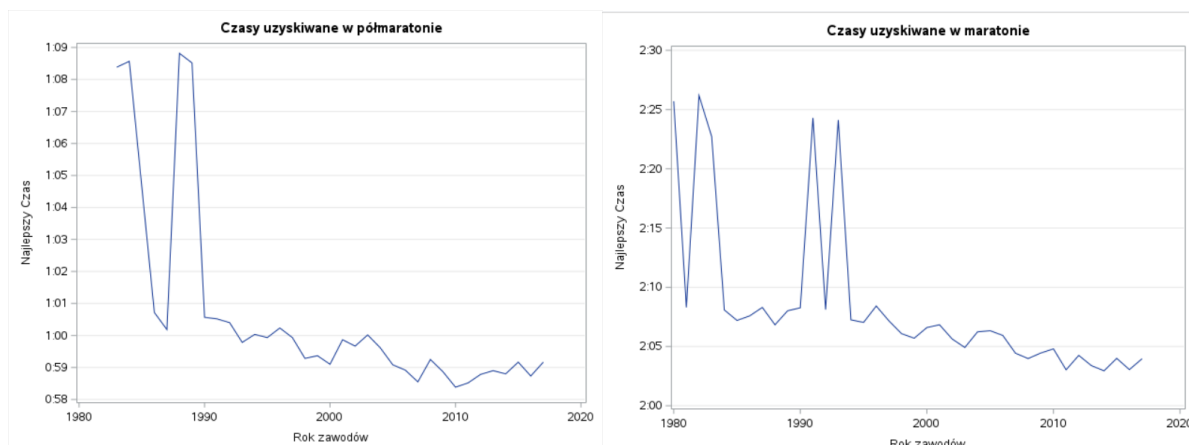
W przypadku dystansów 400, 800 oraz 1500 metrów wykres wygląda bardzo podobnie jak dla 100 i 200 metrów.



Na wykresie prezentującym 5 km widać stopniową poprawę wyników na przestrzeni lat z lekką przerwą, która wystąpiła około 2020. Następnie po osiągnięciu rekordu świata wyniki spadały z gwałtownym spadkiem w 2017 roku. Może on jednak wynikać z okresu pobierania tych danych w tym roku i lepszymi osiągniętymi wynikami w późniejszym okresie 2017 roku.



W przypadku 10 km również wyniki ulegały poprawie na przestrzeni lat. W 2000 roku nastąpił spadek, a kilka lat później rekord świata. Podobnie jak w przypadku 5 km następnie widać spadek wyników.



Analizując wyniki w półmaratonie oraz maratonie można dojść do wniosku, że pomimo wahań w latach 1980 – 1985, poprawa wyników następuje praktycznie cały czas z bardzo niewielkimi odchyłami.

Po przeanalizowaniu wyników można dojść do wniosku, że różnie rozkładały się one w czasie na poszczególnych dystansach. Dla biegów najdłuższych w naszej analizie widać stabilną poprawę i częste ustanawianie rekordów. W przypadku krótszych biegów występują znaczne wahania. Są one widoczne szczególnie w latach 1970-1980. Analiza pokazuje również, że rekordy świata padały zwykle między 2000 a 2010 rokiem. Możliwe, że spadki wyników w późniejszym okresie są spowodowane zaostreniem kontroli antydopingowej. Rozwinęła się ona na tyle, że zawodnikom coraz trudniej jest ją oszukać i staje się coraz bardziej skuteczna. Warto zaznaczyć również, że niektóre zauważone spadki w wynikach w konkretnych latach mogą być wynikiem bieżących wydarzeń, takich jak czynniki społeczno-polityczne, które mogą wpływać na sport na światową skalę.

Analiza osiągniętych wyników ze względu na płeć zawodnika

Płeć zawodnika to jeden z czynników, który ma kluczowy wpływ na wynik zawodnika. Analiza osiągniętych wyników ze względu na płeć pozwala na zrozumienie ewentualnych różnic w osiągniętych rezultatach między mężczyznami a kobietami. Poniżej przedstawiam, co można sprawdzić poprzez taką analizę:

```
proc means data=Projekt.daneg;
  class Płeć DystansM;
  var Czas;
  output out=Projekt.Analiza4I mean=MeanTime ;
run;
proc sort data =Projekt.ANALIZA4I;
by DystansM;
run;
proc transpose data=Projekt.analiza4I out=Projekt.Analiza4II;
  by DystansM;
  id Płeć;
  var MeanTime;
run;
data Projekt.Analiza4III;
  set Projekt.Analiza4II;
  if _NAME_ ne "DystansM" then do;
    Roznica = (Wom/Men) - 1 ;
    Procent = put(Roznica, percent8.3);
    output;
  end;
  keep DystansM Men WOM Procent;
run;

proc print data=Projekt.Analiza4III;
run;
```

Kod nr 9. Sprawdzenie procentowych różnic w średnich wynikach według płci

| Obs. | DystansM | Men | Wom | Procent |
|------|----------|-------------|-------------|---------|
| 1 | 100 | 0:00:09.945 | 0:00:10.937 | 9.975% |
| 2 | 200 | 0:00:20.041 | 0:00:22.244 | 10.99% |
| 3 | 400 | 0:00:44.477 | 0:00:49.895 | 12.18% |
| 4 | 800 | 0:01:43.738 | 0:01:57.438 | 13.21% |
| 5 | 1500 | 0:03:31.758 | 0:04:00.494 | 13.57% |
| 6 | 5000 | 0:13:01.366 | 0:14:49.170 | 13.80% |
| 7 | 10000 | 0:27:22.016 | 0:31:22.397 | 14.64% |
| 8 | 21097 | 1:00:10.097 | 1:08:40.789 | 14.15% |
| 9 | 42195 | 2:07:09.491 | 2:24:19.296 | 13.50% |

Jak pokazują dane procentowa różnica między wynikami kobiet i mężczyzn rośnie wraz z wydłużaniem dystansu. Na najkrótszych dystansach kobiety osiągają wyniki około 10% gorsze od mężczyzn. Następnie ta różnica waha się między 13-15% na dłuższych dystansach i spada lekko na dystansie maratonu do 13,5%. Pokazuje to różnice w wynikach, które dzielą obie płcie. Postanowiłem również sprawdzić, jak wygląda to w przypadku najlepszych wyników. W tym celu wykorzystałem kod nr 9.

```
proc means data=Projekt.daneg;
  class Płeć DystansM;
  var Czas;
  output out=Projekt.Analiza4II min=MinTime ;
run;
proc sort data =Projekt.ANALIZA4II;
by DystansM;
run;
proc transpose data=Projekt.analiza4II out=Projekt.Analiza4III;
  by DystansM;
  id Płeć;
  var MinTime;
run;
data Projekt.Analiza4IIII;
  set Projekt.Analiza4III;
  if _NAME_ ne "DystansM" then do;
    Różnica = (Wom/Men) - 1 ;
    Procent = put(Różnica, percent8.3);
    output;
  end;
  keep DystansM Men Wom Procent;
run;

proc print data=Projekt.Analiza4IIII;
run;
```

Kod nr 10. Sprawdzenie procentowych różnic w najlepszych wynikach według płci

| Obs. | DystansM | Men | Wom | Procent |
|------|----------|-------------|-------------|---------|
| 1 | 100 | 0:00:09.580 | 0:00:10.490 | 9.499% |
| 2 | 200 | 0:00:19.190 | 0:00:21.340 | 11.20% |
| 3 | 400 | 0:00:43.030 | 0:00:47.600 | 10.62% |
| 4 | 800 | 0:01:40.910 | 0:01:53.280 | 12.26% |
| 5 | 1500 | 0:03:26.000 | 0:03:50.070 | 11.68% |
| 6 | 5000 | 0:12:37.350 | 0:14:11.150 | 12.39% |
| 7 | 10000 | 0:26:17.530 | 0:29:17.450 | 11.41% |
| 8 | 21097 | 0:58:23.000 | 1:04:52.000 | 11.10% |
| 9 | 42195 | 2:02:57.000 | 2:15:25.000 | 10.14% |

W przypadku najlepszych wyników różnica na najkrótszych dystansach również wynosi około 10% jednak nie rośnie ona tak jak w przypadku średnich, osiąga maksymalnie 12,39%. W przypadku maratonu wynosi zaledwie 10,14%.

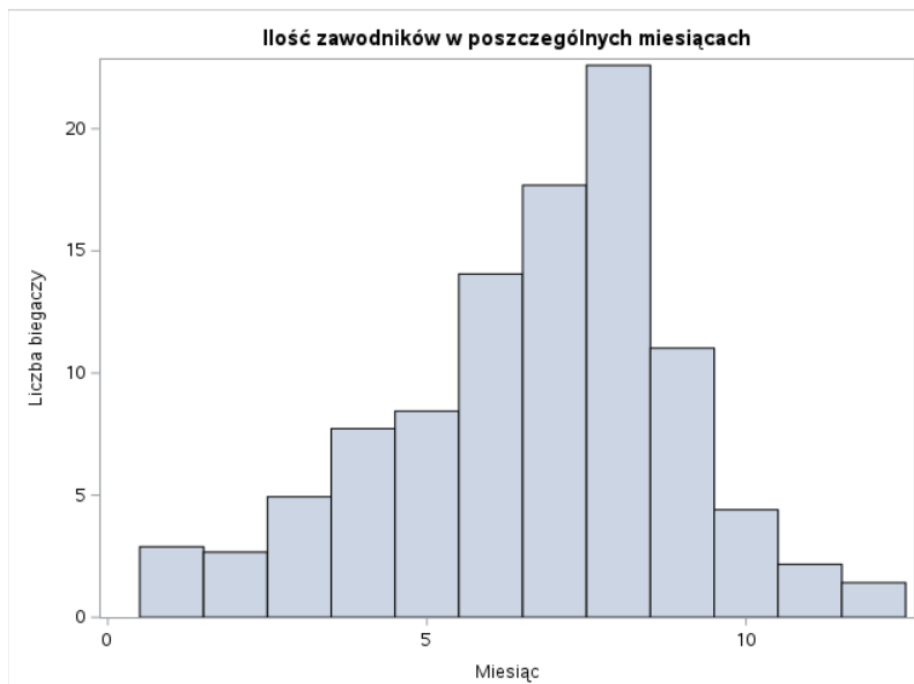
Podsumowując analizę wyników biegaczy ze względu na płeć, można zauważyć interesujące różnice, które kształtują się w zależności od długości dystansu. Analiza średnich czasów pokazuje, że procentowa różnica między wynikami kobiet a mężczyzn rośnie wraz ze wzrostem długości dystansu. Na najkrótszych dystansach, kobiety osiągają wyniki około 10% gorsze niż mężczyźni, a ta różnica wzrasta do 13-15% na dłuższych dystansach, by lekko spaść na dystansie maratonu do 13,5%. Wyniki te potwierdzają istniejące biologiczne różnice między płciami, takie jak różnice w budowie ciała, gęstości mięśni czy pojemności płuc, które mogą wpływać na osiągnane rezultaty. Analiza najlepszych wyników pokazuje pewne zaskakujące tendencje. Pomimo istniejących różnic biologicznych, procentowa różnica między najlepszymi wynikami kobiet a mężczyzn na najkrótszych dystansach wynosi również około 10%, podobnie jak dla średnich wyników. Co ciekawe, różnica ta nie rośnie tak dynamicznie na dłuższych dystansach, osiągając maksymalnie 12,39% na przestrzeni maratonu. To sugeruje, że w przypadku najlepszych zawodniczek różnice te mogą być mniej znaczące, a pewne elementy techniki biegu czy strategii treningowej mogą wpływać na zniwelowanie biologicznych różnic między płciami.

Analiza wpływu daty i miejsca zawodów na uzyskiwane wyniki.

W tej analizie sprawdzę, jak wyglądają uzyskiwane wyniki ze względu na datę i miejsce zawodów. Powinno dać to odpowiedź na pytanie w jakim okresie roku dyspozycja zawodników jest najlepsza oraz czy istnieją miejsca, które ze względu np. na swoją topografię, dają możliwość uzyskania wyższych prędkości podczas biegu.

```
Data Projekt.Analiza5;  
set PROJEKT.daneg;  
Miesiac = Month(Data);  
run;  
  
proc sgplot data=Projekt.Analiza5;  
  histogram Miesiac / binwidth=1;  
  title1 'Ilość zawodników w poszczególnych miesiącach ';  
  xaxis label="Miesiąc" integer discreteorder=data;  
  yaxis label="Liczba biegaczy";  
run;  
  
proc freq data=Projekt.Analiza5;  
  tables Miesiac / nocum nocol nopercnt out=Projekt.Analiza5I;  
  
run;  
  
proc univariate data =Projekt.Analiza5I normal;  
var Miesiac ;  
run;
```

Kod nr 11: Kod do analizy miesiąca zawodów



Jak pokazuje histogram najlepsze wyniki są uzyskiwane na początku drugiej połowy roku. Miesiącem szczególnie obfitym w najlepsze wyniki jest sierpień. Może to być spowodowane dobrymi warunkami atmosferycznymi, jest to ciepły miesiąc, ale chłodniejszy niż np. lipiec, który często jest najcieplejszym miesiącem roku. Histogram wygląda podobnie jak rozkład normalny, więc postanowiłem wykonać test, aby to sprawdzić.

| Testy normalności | | | | |
|----------------------|------------|----------|----------------|---------|
| Testowanie | Statystyka | | Wartość p | |
| Shapiro-Wilka | W | 0.966896 | Pr. < W | 0.8757 |
| Kołmogorowa-Smirnowa | D | 0.089296 | Pr. > D | >0.1500 |
| Cramera-von Misesa | W-kwadr. | 0.019774 | Pr. > W-kwadr. | >0.2500 |
| Andersona-Darlinga | A-kwadr. | 0.153024 | Pr. > A-kwadr. | >0.2500 |

W przypadku dużych prób korzystamy z testu Kołmogorowa-Smirnowa. Wartość p wskazuje, iż faktycznie jest to rozkład normalny. Następnie postanowiłem sprawdzić w których miejscach zawodnicy biegli wyniki, które wystarczyły, aby znaleźć się w TOP1000 na poszczególnych dystansach z uwzględnieniem płci.

```
proc freq data=Projekt.daneg order=freq;
tables Miasto / out=Projekt.Analiza5II (keep=Miasto Count );
run;
```

Kod nr 12: Procedura FREQ w celu ustalenia liczebności.

| Procedura FREQ | | | | |
|----------------|------------|---------|------------------------|---------------------|
| Miasto | Liczebność | Procent | Liczebność skumulowana | Procent skumulowany |
| Zürich | 801 | 4.39 | 801 | 4.39 |
| Bruxelles | 758 | 4.15 | 1559 | 8.55 |
| London | 676 | 3.71 | 2235 | 12.25 |
| Berlin | 674 | 3.69 | 2909 | 15.94 |
| Roma | 649 | 3.56 | 3558 | 19.50 |
| Eugene | 578 | 3.17 | 4136 | 22.67 |
| Monaco | 565 | 3.10 | 4701 | 25.77 |
| Saint-Denis | 442 | 2.42 | 5143 | 28.19 |
| Oslo | 433 | 2.37 | 5576 | 30.56 |
| Beijing | 354 | 1.94 | 5930 | 32.50 |
| Lausanne | 347 | 1.90 | 6277 | 34.41 |
| Athina | 330 | 1.81 | 6607 | 36.21 |
| Moskwa | 314 | 1.72 | 6921 | 37.94 |
| Stockholm | 291 | 1.60 | 7212 | 39.53 |
| Tokyo | 249 | 1.36 | 7461 | 40.90 |
| Hengelo | 244 | 1.34 | 7705 | 42.23 |
| Rieti | 241 | 1.32 | 7946 | 43.55 |
| New York City | 229 | 1.26 | 8175 | 44.81 |
| Paris | 212 | 1.16 | 8387 | 45.97 |
| Kingston | 205 | 1.12 | 8592 | 47.09 |

Jak pokazuje liczebność topowymi miastami są: Zurich, Bruksela, Londyn, Berlin, Rzym. Są to kraje europejskie, co pokazuje, że zawodnicy z całego świata przyjeżdżają do Europy, aby biegać najlepsze wyniki. Jest to związane z dużą liczbą imprez biegowych w tych miastach oraz wysokimi nagrodami za wygranę zawodów lub ustanowienie rekordu trasy.

Przeprowadzona analiza potwierdziła, że są okresy w ciągu roku, kiedy biegacze uzyskują najlepsze wyniki. Jest to początek 2 połowy roku. Ta informacja może być kluczowa, aby zaplanować sezon zawodnika, okresy przygotowawcze, szczyt formy oraz roztrenowanie. Analiza pokazała również, że rozkład ten przypomina rozkład normalny. Aspekt geograficzny pokazał, że najlepsze wyniki są uzyskiwane w Europie, konkretnie w miastach: Zurich, Bruksela, Londyn, Berlin, Rzym. To również jest kluczowa informacja, aby wybrać miejsce zawodów zawodnika.

Wnioski

Bieganie jest jedną z najbardziej dostępnych i popularnych form aktywności fizycznej na całym świecie. Jego prostota sprawia, że może być praktykowane zarówno przez amatorów, jak i profesjonalistów, niezależnie od wieku czy poziomu sprawności. Świat biegania obejmuje różnorodne dyscypliny, takie jak biegi krótkodystansowe, średniodystansowe, maratony. Każda z tych form ma swoje specyficzne cechy, które można analizować. Istnieje wiele parametrów do rozważenia np. czas, dystans, płeć. Analiza biegania staje się coraz bardziej zaawansowanym narzędziem wspierającym rozwój tego sportu. Przeprowadzona analiza różnych aspektów wyników biegowych dostarcza wnikliwego spojrzenia na złożoność czynników wpływających na osiągnięcia zawodników. Poniżej przedstawiam podsumowanie głównych wniosków wynikających z przeprowadzonych analiz.

Analiza osiągniętych wyników ze względu na narodowość zawodnika:

Analiza ta ukazuje, że są kraje, w których zawodnicy częściej uzyskują lepsze wyniki na poszczególnych dystansach niż reprezentanci innych krajów. Może to być związane zarówno z systemem szkoleniowym, warunkami atmosferycznymi czy genetyką.

Analiza osiągniętych wyników ze względu na płeć zawodnika:

Analiza z podziałem na płeć ukazuje istotne różnice w osiągniętych wynikach między mężczyznami a kobietami. Odkryto, że zarówno średnie czasy biegów, jak i te najlepsze różnią się w zależności od płci.

Analiza osiągniętych wyników ze względu na wiek zawodnika:

Analiza dotyczyła wieku, w którym zawodnicy osiągają swoje szczytowe wyniki. Wykresy wiekowe wykazały, że zdecydowana większość sportowców, bez względu na płeć, osiąga swoje najlepsze wyniki biegowe między 20 a 30 rokiem życia. Te wyniki sugerują, że wiek ten jest kluczowy dla szczytu wydolności zawodników. Dane wskazują, iż zwykle kobiety osiągają ten szczyt nieco później.

Analiza osiąganych wyników w czasie (historycznie):

Analiza czasowa pokazała, że kluczowy w tym aspekcie jest dystans. Na biegach długich progres wydaje się stały, gdy na krótszych często się on wacha. Wykazano również, że był okres, kiedy rekordy na wielu dystansach były ustanawiane, a następnie nastąpił spadek wyników.

Analiza wpływu daty i miejsca zawodów na uzyskiwane wyniki:

Analiza ta wykazała, że są okresy w sezonie, kiedy zawodnicy biegają najlepiej i to właśnie na nich powinniśmy się skupić, przygotowując plan na sezon. Dodatkowo istnieją miasta, które ze względu na wiele aspektów np. topografię, ilość zawodów sprzyjają uzyskiwaniu dobrych czasów przez zawodników.

Podsumowując, przeprowadzona analiza dostarcza kompleksowego spojrzenia na wielowymiarowe aspekty wyników biegowych. Zrozumienie wpływu płci, wieku, narodowości, czasu i miejsca na osiągnięte rezultaty stanowi cenny wkład w lepsze kierowanie treningiem, oraz opracowywanie skuteczniejszych strategii przygotowywania zawodników na zawody.