Projekt nr 2

Marcin Kapłon

16.01.2022r.

Graficzna analiza danych o Igrzyskach Olimpijskich

Wczytanie danych i krótki opis

```
igrzyska<-read.csv("olympics.csv")
library(dplyr)
library(tidyr)
library(ggplot2)
library(RColorBrewer)
head(igrzyska, 6)
```

```
##   ID                    Name Sex Age Height Weight          Team NOC
## 1  1               A Dijiang   M  24    180     80         China CHN
## 2  2                A Lamusi   M  23    170     60         China CHN
## 3  3     Gunnar Nielsen Aaby   M  24     NA     NA        Denmark DEN
## 4  4    Edgar Lindenau Aabye   M  34     NA     NA Denmark/Sweden DEN
## 5  5 Christine Jacoba Aaftink   F  21    185     82    Netherlands NED
## 6  5 Christine Jacoba Aaftink   F  21    185     82    Netherlands NED
##         Games Year Season      City          Sport
## 1 1992 Summer 1992 Summer Barcelona     Basketball
## 2 2012 Summer 2012 Summer    London          Judo
## 3 1920 Summer 1920 Summer Antwerpen       Football
## 4 1900 Summer 1900 Summer     Paris     Tug-Of-War
## 5 1988 Winter 1988 Winter   Calgary Speed Skating
## 6 1988 Winter 1988 Winter   Calgary Speed Skating
##                            Event Medal
## 1       Basketball Men's Basketball  <NA>
## 2       Judo Men's Extra-Lightweight  <NA>
## 3           Football Men's Football  <NA>
## 4       Tug-Of-War Men's Tug-Of-War  Gold
## 5   Speed Skating Women's 500 metres  <NA>
## 6 Speed Skating Women's 1,000 metres  <NA>
```

W swojej analizie wykorzystałem plik zawierający dane o uczestnikach Igrzysk Olimpijskich w latach 1896-2016. Każda obserwacja to jeden uczestnik podczas jednej konkurencji w jednym roku. Ramka zawiera informacje o płci, wieku, wzroście i wadze uczestników, o tym jaki kraj reprezentują, w jakiej konkurencji brali udział oraz czy zdobyli medal. Kompletność danych jest bardzo duża, ale nie stuprocentowa, więc mogą istnieć drobne rozbieżności między wynikami a stanem faktycznym.

Plik zawiera informacje zarówno o letnich, jak i zimowych igrzyskach. Ja w swojej analizie skupię się przede wszystkich na tych letnich. Nie będę również uwzględniał danych o Olimpiadzie Letniej z 1906 roku, która zwyczajowo nie jest wliczana do numeracji igrzysk, gdyż była wydarzeniem specjalnym z okazji 10-lecia pierwszych zawodów.

Zastosowane w pracy skróty *IO* i *MKOl* oznaczają kolejno *Igrzyska olimpijskie* i *Międzynarodowy Komitet Olimpijski*.

1. Liczba medali we wszystkich konkurencjach na letnich IO

Stworzę wykres kolumnowy dla 10 państw, które zdobyły najwięcej medali we wszystkich konkurencjach i wszystkich igrzyskach letnich łącznie. Warto uwzględnić, że niektóre konkurencje (np. piłka nożna) są grupowe, a nie możemy przypisać jednemu państwu wielu medali za to samo osiągnięcie. Na przykład jeśli Brazylia wygrała turniej piłki nożnej mężczyzn w 2016 roku, to obecnie w tabeli przy Brazylii jest kilkanaście złotych medali za to samo. Aby rozwiązać ten problem, musimy wybrać odpowiednie zmienne i użyć funkcji `unique()`.

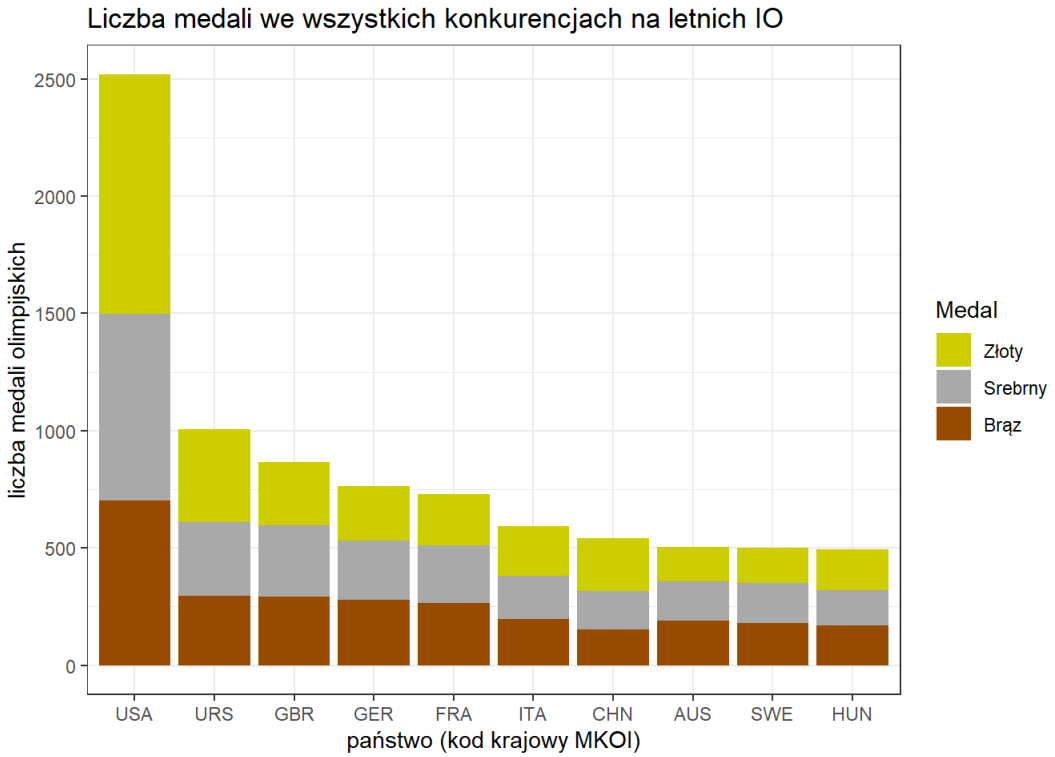
```
klasyfikacja<-igrzyska%>%
  filter(grepl("Summer",Games), Year!=1906)%>%  #uwzględnienie tylko letnich
  select(NOC, Medal, Event, Year)%>%
  unique()%>%       #eliminuję powtarzające sie medale
  filter(!is.na(Medal))%>%
  group_by(NOC)%>%
  count(Medal)%>%
  spread(Medal,n)%>%  #używam spread aby dodać sumę medali
  arrange(desc(Gold))%>%
  group_by(NOC)%>%
  mutate(suma=Gold+Silver+Bronze)%>%
  gather(Medal, n,-NOC, -suma)%>%  #powracam do stanu sprzed użycia spread()
  arrange(desc(suma))%>%  #teraz gdy mam kolumnę 'suma' mogę po niej sortować
  mutate(Medal=factor(Medal, levels=c("Gold","Silver","Bronze"))) #zmieniam kolejnosc danych na wykresie (faktoryzacja)

klasyfikacja
```

```
## # A tibble: 441 x 4
## # Groups:   NOC [147]
##    NOC    suma Medal      n
##    <chr> <int> <fct>  <int>
##  1 USA    2521 Bronze   701
##  2 USA    2521 Gold    1023
##  3 USA    2521 Silver   797
##  4 URS    1005 Bronze   294
##  5 URS    1005 Gold     394
##  6 URS    1005 Silver   317
##  7 GBR     867 Bronze   293
##  8 GBR     867 Gold     270
##  9 GBR     867 Silver   304
## 10 GER     762 Bronze   278
## # ... with 431 more rows
```

Teraz stworzę wykres na podstawie otrzymanej tabeli.

```
ggplot(klasyfikacja[1:30,])+
geom_col(aes(x=reorder(NOC, -n),
             y=n,
             fill=Medal))+
labs(title="Liczba medali we wszystkich konkurencjach na letnich IO",
       x="państwo (kod krajowy MKOl)",
       y="liczba medali olimpijskich")+
scale_fill_manual(values=c("yellow3","darkgrey","#964B00"),
                  labels=c("Złoty","Srebrny","Brąz"))+
theme_bw()
```



Na podstawie wykresu możemy wyciągnąć kilka ciekawych wniosków. Stany Zjednoczone nie tylko zdobyły najwięcej medali olimpijskich, ale też mają ich aż o około 2,5 raza więcej od nieistniejącego już Związku Radzieckiego, który w otrzymanym rankingu znajduje się na drugim miejscu. Liczba samych złotych medali USA jest zbliżona do liczby medali ZSRR ogółem. Warto mieć na uwadze, że Związek Radziecki istniał o wiele krócej niż USA, które od 1896 roku zachowują ciągłość historyczną, stąd ten wynik należy potraktować tylko jako ciekawostkę.

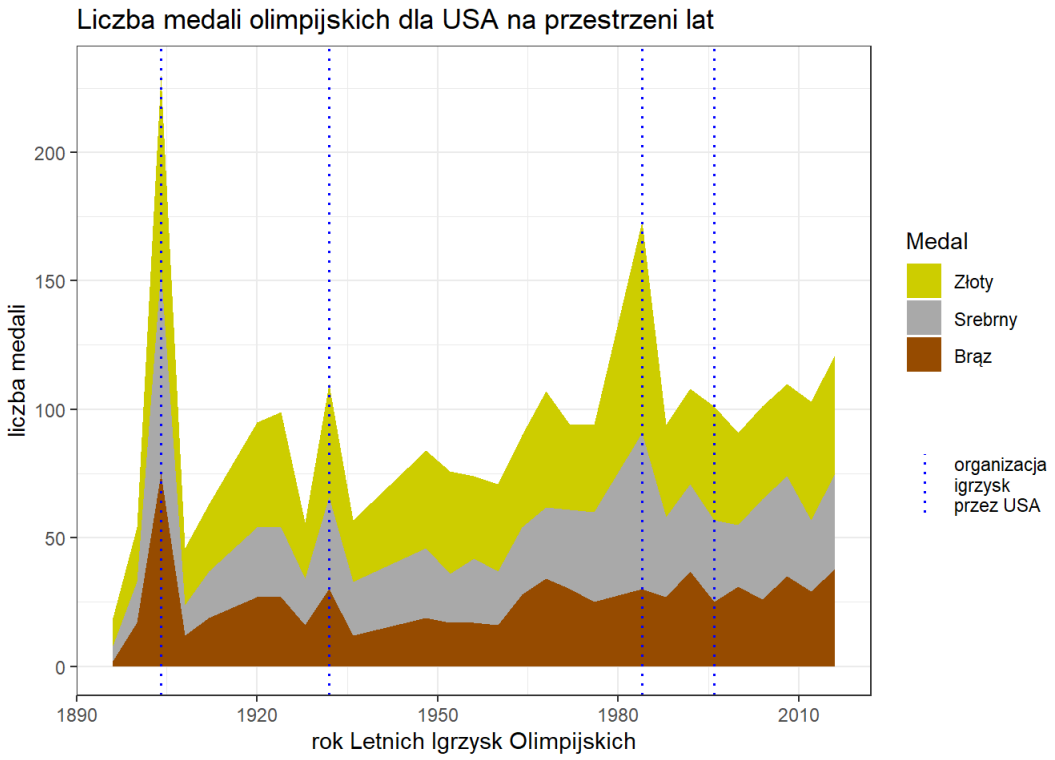
Inną ciekawą obserwacją jest to, że spośród tych 10 państw aż 7 z nich leży w Europie. Brakuje za to wielu państw z dużą populacją takich jak Indie. Zaskoczeniem może być też wysoki wynik Węgier, w których mieszka około 10 milionów ludzi. Można wyciągnąć wniosek, że bycie małym krajem nie stoi w sprzeczności z osiąganiem dobrych wyników na igrzyskach.

Spośród wszystkich medali Stanów Zjednoczonych wyraźnie najwięcej z nich jest złotych. Podobną zależność widać w przypadku Chin i ZSRR. Z kolei inne kraje mają podobną liczbę medali każdego rodzaju.

2. Liczba medali olimpijskich dla USA na przestrzeni lat

Wysoki wynik w ogólnej klasyfikacji medalowej USA skłonił mnie do zbadania liczby medali dla Stanów Zjednoczonych na przetrzeni lat. Ponownie wezmę pod uwagę jedynie Letnie Igrzyska Olimpijskie. Wykorzystam wykres warstwowy.

```
igrzyska%>%
  filter(NOC=="USA", grepl("Summer",Games), Year!=1906)%>%
  select(Medal, Event, Year)%>%
  unique()%>%
  filter(!is.na(Medal))%>%
  count(Year,Medal)%>%
  mutate(Medal=factor(Medal, levels=c("Gold","Silver","Bronze")))%>%
  ggplot()+
  geom_area(aes(x=Year,
                y=n,
                fill=Medal), position="stack")+
  scale_fill_manual(values=c("yellow3","darkgrey","#964B00"),
                    labels=c("Złoty", "Srebrny", "Brąz"))+
  geom_vline(aes(xintercept = 1904,
                 col="organizacja\nigrzysk\nprzez USA"),
                 linetype="dotted", size=0.6)+
  geom_vline(aes(xintercept = 1932,
                 col="organizacja\nigrzysk\nprzez USA"),
                 linetype="dotted", size=0.6)+
  geom_vline(aes(xintercept = 1984,
                 col="organizacja\nigrzysk\nprzez USA"),
                 linetype="dotted", size=0.6)+
  geom_vline(aes(xintercept = 1996,
                 col="organizacja\nigrzysk\nprzez USA"),
                 linetype="dotted", size=0.6)+
  scale_color_manual(name="",values=c("organizacja\nigrzysk\nprzez USA"="blue"))+
  labs(title="Liczba medali olimpijskich dla USA na przestrzeni lat",
       x="rok Letnich Igrzysk Olimpijskich",
       y="liczba medali")+
  theme_bw()
```



Podczas tworzenia pierwszej wersji tego wykresu (bez przerywanych linii) zauważyłem nagłe bardzo duże wzrosty liczby medali w niektórych latach, w tym jeden szczególnie duży w 1904 roku. Stwierdziłem, że mogło być to spowodowane na przykład zmianą liczby konkurencji albo wystawieniem rekordowo dużej liczby zawodników (Obie te hipotezy zweryfkuję później). Zauważyłem jednak również, że w 1904 roku organizatorem igrzysk były USA.

```
igrzyska%>%
  filter(Year==1904)%>%
  select(City)%>%
  unique()
```

```
##       City
## 1 St. Louis
```

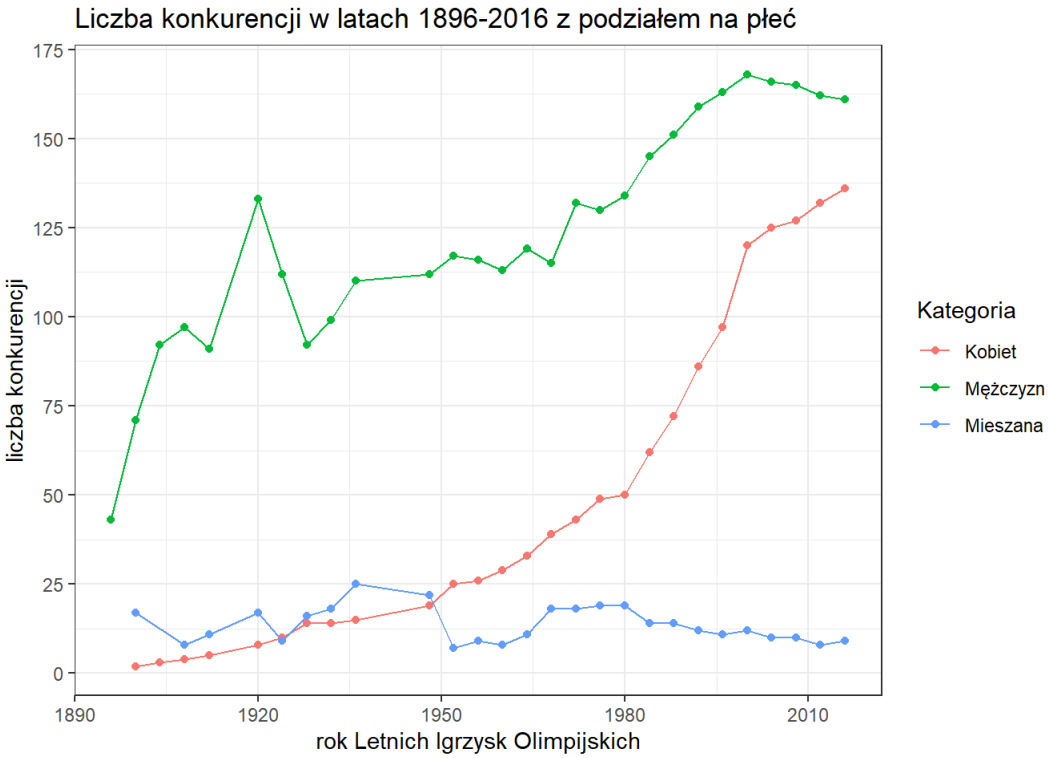
Po naniesieniu na wykres informacji o latach organizowania igrzysk przez USA okazało się, że w latach, w których Stany Zjednoczone były gospodarzem, zdobywały również rekordowo dużo medali. Być może więcej zawodników było w stanie wtedy wystartować, gdyż igrzyska były bliżej oraz pewnie bardziej rozpromowane. Co ciekawe, wzrost liczby medali nie nastąpił w 1996 roku, co może świadczyć np. o późniejszym wprowadzeniu przez Komitet Olimpijski jakichś regulacji zapobiegającym sytuacjom, gdzie gospodarzom łatwiej wygrywać igrzyska.

Z wykresu możemy odczytać również, że jeśli nie uwzględnimy tych nagłych wzrostów, to w USA jest tendencja wzrostowa w zdobywaniu medali. Dane są liczbowe, a nie procentowe, więc niekoniecznie oznacza to coraz lepsze wyniki w klasyfikacji medalowej, a może być powiązane przykładowo ze zmianą liczby konkurencji.

3. Liczba konkurencji w latach 1896-2016 z podziałem na płeć

Myślę, że warto zbadać jak zmieniała się liczba konkurencji w Letnich Igrzyskach Olimpijskich z kilku powodów. Po pierwsze, sprawdzimy czy pokrywa się ona z liczbą medali dla USA, a także dowiemy się wstępnie jak zmieniały się dysproporcje płciowe na igrzyskach. Przy tworzeniu wykresu trzeba uwzględnić to, że na igrzyskach odbywają się równiez konkurencje mieszane, w których biorą udział zarówno mężczyźni, jak i kobiety.

```
igrzyska%>%
  mutate(kategoria=ifelse(
    grepl("Women's", Event),"Kobiet",ifelse(
      grepl("Men's",Event),"Mężczyzn","Mieszana")))%>%
  filter(grepl("Summer",Games), Year!=1906)%>%
  select(Year, Event, kategoria)%>%
  unique()%>%
  count(Year, kategoria)%>%
  ggplot(aes(x=Year))+
  geom_line(aes(y=n, col=kategoria))+
  geom_point(aes(y=n, col=kategoria))+
  labs(title="Liczba konkurencji w latach 1896-2016 z podziałem na płeć",
       x="rok Letnich Igrzysk Olimpijskich",
       y="liczba konkurencji",
       color="Kategoria")+
  scale_y_continuous(breaks=seq(0,175,25))+
  theme_bw()
```



Zdaje się, że liczba konkurencji na igrzyskach olimpijskich w lecie jest powiązana z rosnącą tendencją zdobywania medali przez USA. Dodatkowo, wzrost liczby konkurencji jest o wiele szybszy niż przyrost liczby medali na wcześniejszym wykresie, więc możemy wnioskować, że USA z igrzysk na igrzyska mają procentowo coraz mniejszy udział we wszystkich medalach. Za pomocą powyższego wykresu możemy wytłumaczyć również wzrost liczby medali dla Stanów Zjednoczonych w 1920 roku, którego nie dało się powiązać z organizacją igrzysk. Liczba konkurencji w tym czasie bowiem znacząco wzrosła. Szczególnie ciekawą informacją jest to, że w 1904 roku było trochę mniej niż 100 konkurencji, a USA zdobyły ponad 225 medali, co oznacza, że w dużej części konkurencji musiały zająć wszystkie 3 miejsca.

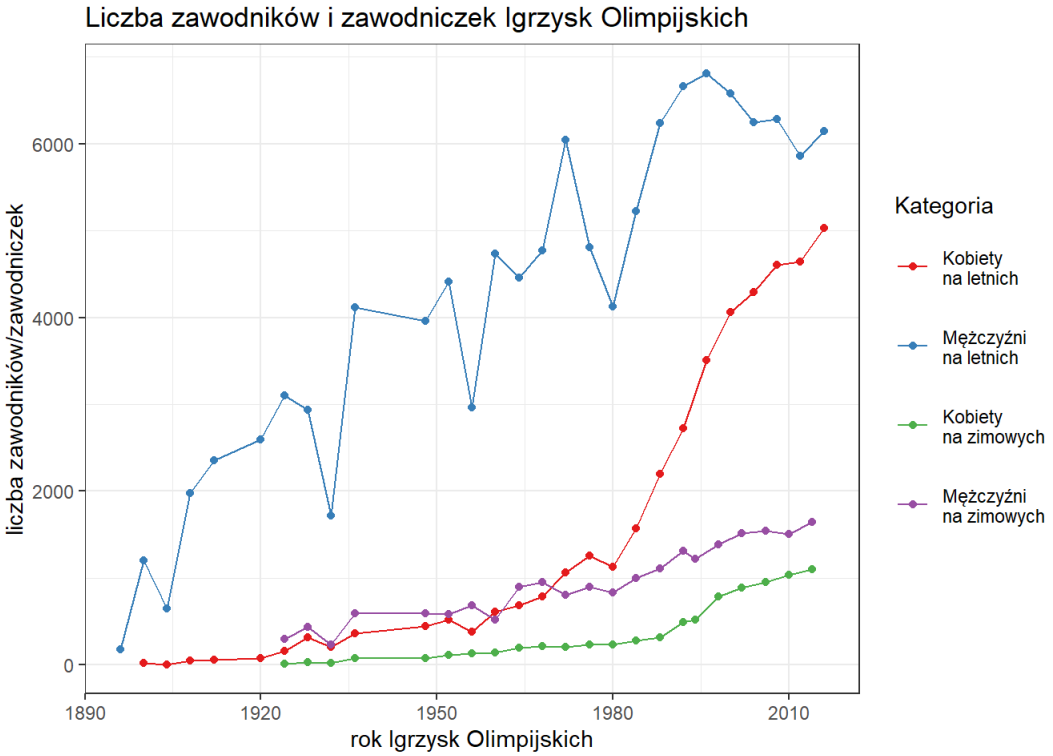
Jednak przy analizie powyższego wykresu o wiele bardziej inetersującą kwestią jest duża dysproporcja pomiędzy liczbą męskich i żeńskich konkurencji. Łatwo zauważyć, że na pierwszych igrzyskach w 1896 roku w zawodach mogli brać udział jedynie mężczyźni. Pierwsze damskie i mieszane konkurencje pojawiły się już na kolejnych igrzyskach, ale ich liczba była symboliczna. Z każdymi kolejnymi zawodami pojawiało się systematycznie coraz więcej kobiecych konkurencji, aż w 1980 roku liczba ta zaczęła rosnąć znacznie szybciej. Możemy podejrzewać, że w kolejnych latach liczba ta wyrówna się z konkurencjami męskimi.

Z kolei, jeśli chodzi o męskie konkurencje, to od 2000 roku widać tendencję spadkową. Podobnie od 1980 roku zaczęła powoli maleć liczba konkurencji mieszanych, która zresztą nigdy nie była większa od 25. Ogólnie, liczba konkurencji olimpijskich wzrosła od pierwszych zawodów do 2016 roku z około 40 do ponad 300 i można się domyślać, że utrzyma się na tym poziomie, ale z coraz mniejszą dysproporcją między kobiecymi a męskimi.

4. Liczba zawodników Igrzysk Olimpijskich na przestrzeni lat

Pomimo tego, że znamy już liczbę konkurencji olimpijskich na przestrzeni lat, to warto zbadać dodatkowo dla pełniejszego obrazu sytuacji, ile było zawodników i zawodniczek IO. Tym razem, dla urozmaicenia wykresu, przedstawię również dane dla Zimowych Igrzysk Olimpijskich.

```
igrzyska%>%
  filter(Year!=1906)%>%
  select(Year, Season, Sex, Name)%>%
  unique()%>%
  count(Year, Season, Sex)%>%
  unite("Season_Sex", Season, Sex)%>%
  ggplot(aes(x=Year, y=n, col=Season_Sex))+
  geom_line(size=.5)+
  geom_point()+
  labs(title="Liczba zawodników i zawodniczek Igrzysk Olimpijskich",
       x="rok Igrzysk Olimpijskich",
       y="liczba zawodników/zawodniczek",
       color="Kategoria")+
  scale_color_brewer(labels=c("\nKobiety\nna letnich\n","\nMężczyźni\nna letnich\n", "\nKobiety\nna zimowych\n", "\nMężczyźni\nna zimowych\n"),
                     palette="Set1")+
  theme_bw()
```

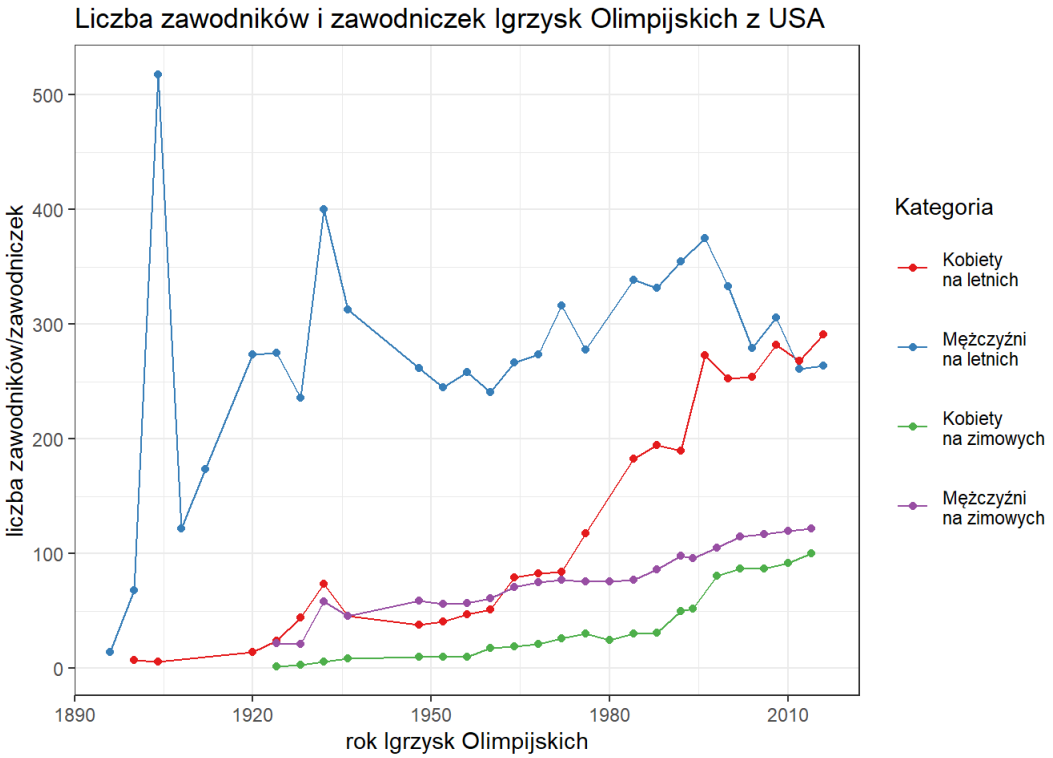


Na tym wykresie zmiany są już o wiele większe. O ile w pierwszych IO brało udział mniej niż 250 osób, to w 2016 roku było to już ponad 11 tysięcy (mowa oczywiście o igrzyskach letnich). Często również występowały duże wahania w liczbie zawodników płci męskiej. Podobnie jak w przypadku liczby konkurencji kobiecych, liczba zawodniczek zaczęła znacząco rosnąć po 1980 roku, choć tutaj wydaje się to być jeszcze wyraźniejsze.

W przypadku Zimowych Igrzysk Olimpijskich możemy zauważyć kilka ciekawych faktów. Pierwsze igrzyska odbyły się w 1924 roku. Z łatwością można dostrzec, że do pewnego momentu 4 punkty na wykresie leżą na jednej linii pionowej, co oznacza, że przez długi czas letnie i zimowe igrzyska odbywały się w tym samym roku. Jeśli chodzi o samą liczbę zawodników, to jest ona o wiele mniejsza od liczby zawodników na igrzyskach letnich, ale wahania losowe nie są tak liczne. Liczba zawodników stale rośnie, ale bardzo wolno. Mężczyzn jest o wiele więcej niż kobiet (i to stale mniej więcej o 300 osób) i nie widać, aby ta dysproporcja się wyrównywała.

Sprawdzę jeszcze, jak ta statystyka wygląda dla samych Stanów Zjednoczonych, aby wreszcie sprawdzić, czy słusznym jest przypuszczenie, że USA miały najwięcej medali, gdy organizowały igrzyska, dlatego że wystawiły wtedy dużo zawodników.

```
igrzyska%>%
  filter(Year!=1906, NOC=="USA")%>%
  select(Year, Season, Sex, Name)%>%
  unique()%>%
  count(Year, Season, Sex)%>%
  unite("Season_Sex", Season, Sex)%>%
  ggplot(aes(x=Year, y=n, col=Season_Sex))+
  geom_line(size=.5)+
  geom_point()+
  labs(title="Liczba zawodników i zawodniczek Igrzysk Olimpijskich z USA",
       x="rok Igrzysk Olimpijskich",
       y="liczba zawodników/zawodniczek",
       color="Kategoria")+
  scale_color_brewer(labels=c("\nKobiety\nna letnich\n","\nMężczyźni\nna letnich\n", "\nKobiety\nna zimowych\n", "\nMężczyźni\nna zimowych\n"),
                     palette="Set1")+
  theme_bw()
```



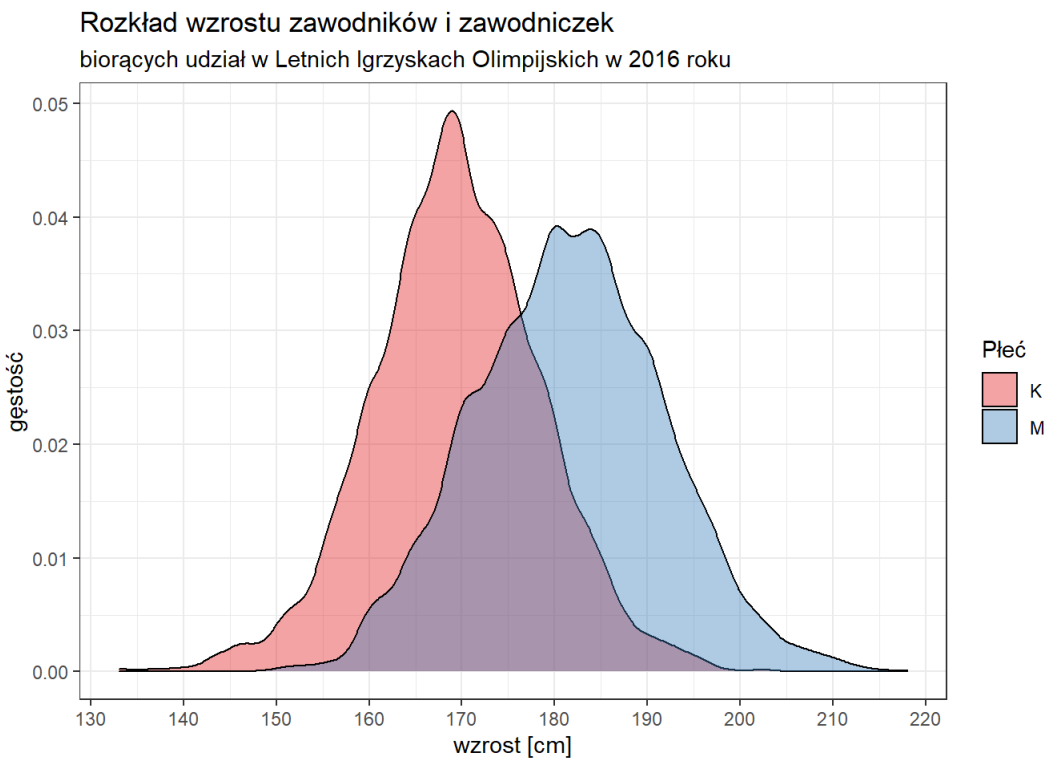
Jak widać Stany Zjednoczone rzeczywiście wystawiały rekordowo dużo zawodników w 1904 i 1932 roku. Dodatkowo z wcześniejszego wykresu możemy odczytać, że w tych latach nastąpiły duże spadki liczby zawodników ogółem przy jednoczesnych dużych wzrostach liczby zawodników z USA, co spowodowało, że w 1904 roku większość wszystkich zawodników to byli Amerykanie, a w 1932 roku stanowili oni co najmniej jedną piątą wszystkich uczestników. Natomiast w 1984 roku (kiedy USA organizowały igrzyska letnie i również zdobyły bardzo dużo medali) nie wystąpiła podobna sytuacja. Z wykresu można również odczytać, że od 2012 roku w Letnich Igrzyskach Olimpijskich bierze udział więcej Amerykanek niż Amerykanów.

5. Rozkład wzrostu zawodników i zawodniczek

Postanowiłem zbadać również jak prezentuje się rozkład wzrostu wśród zawodników i zawodniczek IO z 2016 roku. Wziąłem pod uwagę jedynie ostatnie igrzyska, z tych, o których posiadam dane, gdyż podejrzewam, że średni wzrost u ludzi zmienił się przez ostatnie 100 lat, co odbierałoby sens zestawieniu wszech czasów.

```
igrzyska%>%
  filter(Year==2016)%>%
  ggplot()+
  geom_density(aes(x=Height,
                   fill=as.factor(Sex)),
               position="identity",
               alpha=.4,
               col="black")+
  labs(title="Rozkład wzrostu zawodników i zawodniczek",
       subtitle="biorących udział w Letnich Igrzyskach Olimpijskich w 2016 roku",
       x="wzrost [cm]",
       y="gęstość",
       fill="Płeć")+
  scale_fill_brewer(palette = "Set1",
                    label=c("K","M"))+
  scale_x_continuous(breaks=seq(110, 280, 10))+
  theme_bw()
```

```
## Warning: Removed 176 rows containing non-finite values (stat_density).
```



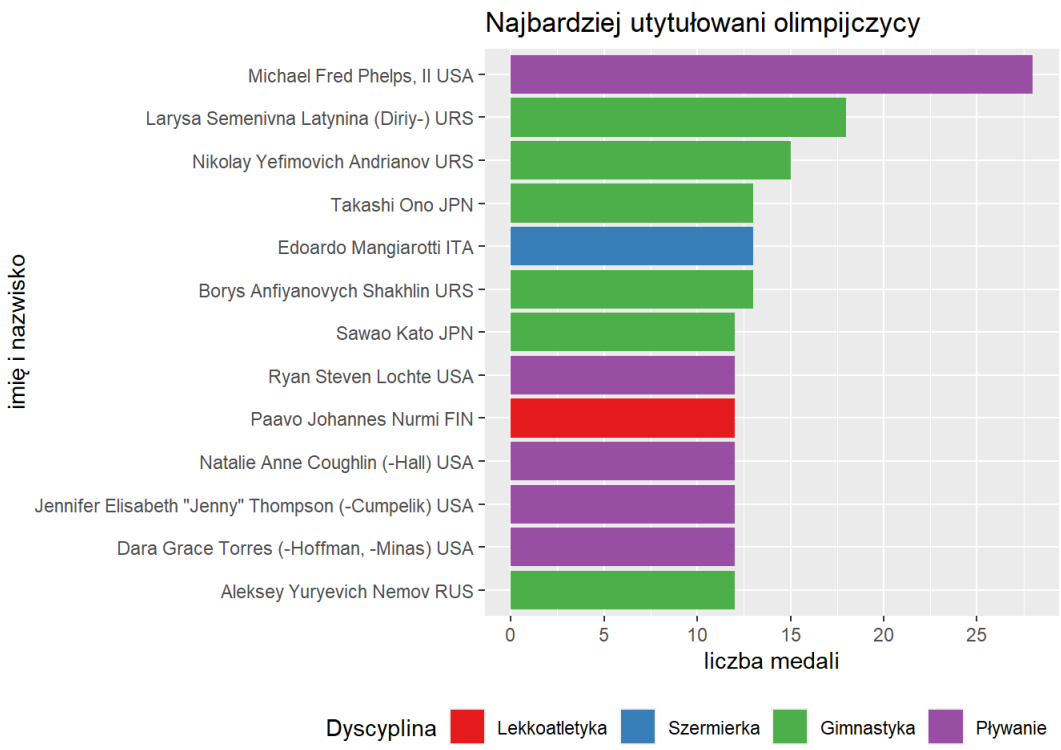
Z danych krzywych gęstości jesteśmy w stanie dowiedzieć się kilku interesujących faktów. Między innymi to, że wśród zawodników płci męskiej krzywa gęstości ma 2 wierzchołki, co oznacza, że rozkład tej zmiennej jest poniekąd dwumodalny (choć ledwo). Taka sytuacja zachodzi często, gdy badana populacja jest połączeniem dwóch innych, co w tej sytuacji jest bardzo możliwe. Mamy do czynienia z zawodnikami wielu różnych dyscyplin, zarówno takich, gdzie wysoki wzrost pomaga osiągać lepsze wyniki, jak i takich, gdzie nie ma to większego wpływu na sukces. Najwięcej zawodników ma około 180 lub 184 cm wzrostu, a najwięcej zawodniczek około 167cm wzrostu.

Co ciekawe funkcja gęstości wśród zawodniczek ma jedynie jedno maksimum lokalne i nie zachodzi tu dwumodalność w żadnym stopniu. Może to być spowodowane tym, że kobiety są do siebie bardziej zbliżone wzrostem, a bimodalność rozkładu i tak była u mężczyzn niewielka. Odchylenie standardowe w rozkładzie wzrostu zawodników płci męskiej jest większe niż u kobiet. Widzimy też, że nie ma prawie żadnej zawodniczki przekraczającej 2m wysokości oraz prawie żadnego sportowca płci męskiej poniżej 1,5m wysokości.

6. Najbardziej utytułowani olimpijczycy

Na samym początku mojej analizy stworzyłem ranking 10 państw, które zdobyły najwięcej medali we wszystkich dyscyplinach razem w historii. Myślę, że ciekawe byłoby poznanie również najbardziej utytułowanych zawodników Letnich Igrzysk Olimpijskich. Wykorzystam do tego celu wykres słupkowy.

```
igrzyska%>%
  filter(grepl("Summer",Games), Year!=1906, !is.na(Medal))%>%
  select(Name, Sport, Medal, NOC)%>%
  count(Name, Sport, NOC)%>%
  arrange(desc(n))%>%
  top_n(10,n)%>%
  mutate(Name=paste(Name,NOC))%>%
  ggplot()+
  geom_col(aes(x=reorder(Name,n), y=n, fill=Sport))+
  coord_flip()+
  theme(legend.position ="bottom",
        legend.justification = "right")+
  labs(title="Najbardziej utytułowani olimpijczycy",
       y="liczba medali",
       x="imię i nazwisko",
       fill="Dyscyplina")+
  scale_fill_brewer(palette="Set1",
    labels=c("Lekkoatletyka","Szermierka","Gimnastyka","Pływanie"))+
  scale_y_continuous(breaks=seq(0, 28, 5))
```



Pochodzący ze Stanów Zjednoczonych Michael Phelps zdobył najwięcej (bo aż 28) medali olimpijskich różnego rodzaju, podczas gdy Larysa Latynina z ZSRR, będąca na drugim miejscu w zestawieniu, szesnaście. Interesujący jest fakt, że pomiędzy osobą na pierwszym i drugim miejscu zestawienia jest tak duża różnica. Co ciekawe, wśród 13 najbardziej utytułowanych olimpijczyków aż 5 było z USA, a 4 z ZSRR/Rosji, co świadczy o tym, że te państwa mogą pochwalić się nie tylko wieloma medalami, ale również najbardziej utytułowanymi zawodnikami. Wśród zawodników spoza 10 państw mających najwięcej medali jest tylko Fin Paavo Nurmi.

Oczywiście wnioskowanie, że Michael Phelps zdobywał te medale w 28 różnych latach byłoby całkowicie chybione, gdyż trzeba pamiętać, że dyscypliny sportowe na IO mogą składać się z wielu konkurencji, co olimpijczycy mogą wykorzystać, by zdobyć więcej niż 1 medal na jednych zawodach. Trzynastu przedstawionych tu zawodników uprawia łącznie 4 dyscypliny sportowe, głównie gimnastykę i pływanie, a wszystkie te dyscypliny składają się z wielu konkurencji. Jak widać, gimnastyka i pływanie najbardziej sprzyjają zdobywaniu wielu medali na IO.

Podsumowanie analizy

Stany Zjednoczone posiadają najwięcej medali olimpijskich (na igrzyskach letnich) oraz mogą pochwalić się najbardziej utytułowanym medalistą, którym jest Michael Phelps. Na Letnich Igrzyskach Olimpijskich pojawiło się ponad 260 nowych konkurencji od pierwszych zawodów, a liczba tych kobiecych stale rośnie i zbliża się do liczby konkurencji męskich. W ostatnich zawodach (dla których są dane) brało udział ponad 11 tysięcy zawodników, z których większość to byli mężczyźni. Dominanta oraz odchylenie standardowe wzrostu uczestników płci męskiej są większe niż u zawodniczek płci żeńskiej.