Projekt

Cezary Moskal

2023-11-15

Wstęp

Cel Badania

Poniższe badanie statytyczne ma na celu sprawdzenie poprawności następujących hipotez:

1. Zależność pomiędzy **urbanizacją**, a **śmiertelnością dziecięcą** nie jest jednoznaczna;
2. Kobiety rodzą więcej dzieci, gdy ludność mieszka przede wszystkim na wsi;
3. Wysoka **gęstość zaludnienia** często idzie razem z rozwojem nierówności społecznych;
4. Wysoka **gęstość zaludnienia** idzie razem z wysoką śmiertelnością dzieci;
5. Relacja między **średnim realnym przychodem**, a nierównościami społecznymi jest zmienna w zależności od pierwszej z tych zmiennych;
6. 20% najbogatszych pod względem średniego mieszkańca krajów posiada 80% całego bogactwa cywilizacyjnego;
7. Zazwyczaj większej **oczekiwanej długości życia** towarzyszy mniejsza **dzietność**.
8. Kraje o większej **powierzchni** posiadają średnio większy odsetek ludzi starszych.

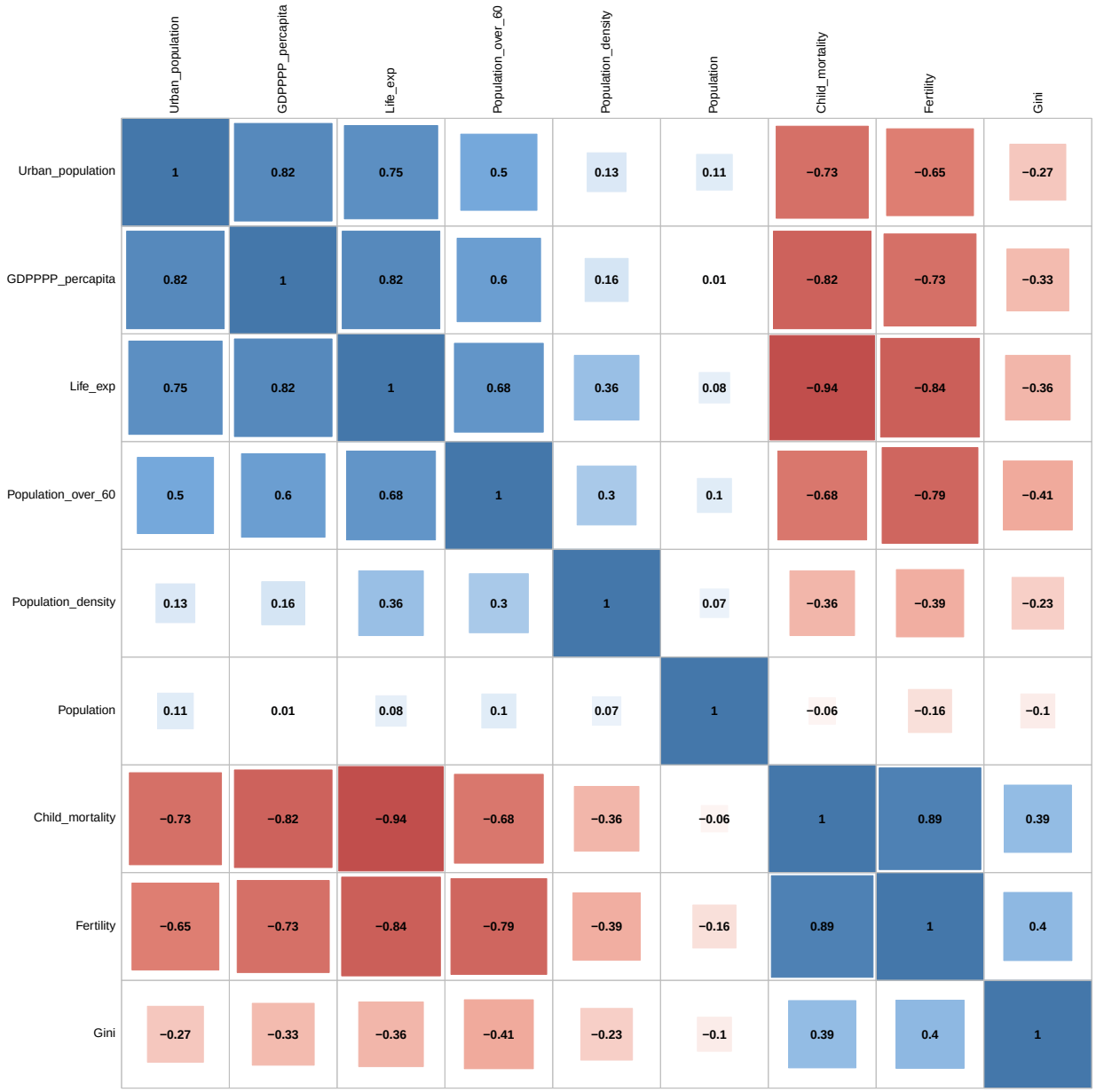
Macierz Korelacji

W celu zobrazowania korelacji między zmiennymi tworzymy macierz korelacji.

Stanowić będzie ona podstawę co do sprawdzenia poprawności wystawionych hipotez.

```
col <- colorRampPalette(
  c('#BB4444', '#EE9988', '#FFFFFF', '#77AADD', '#4477AA'))
correlation <- round(cor(data[3:11], method = 'spearman'), 2)

corrplot(correlation,
         method = 'square', shade.col = NA, tl.col = 'black', tl.srt = 90,
         col = col(100), addCoef.col = 'black', cl.pos = 'n', order = 'AOE')
```




```r
rm(correlation, col)
```

Kilka Wybranych Lat

Wyodrębnijmy lata 1960, 1980, 2000, 2015. Będę one przydatne, gdy chcemy zwizualizować zmienność w czasie w sytuacjach w których przedstawienie wszystkich lat na raz mija się z celem, np. na wykresie liniowym. Dodajmy także zmienną **GDP**, która będzie przedstawiała realne PKB danego kraju w danym roku, a także zmienną **Area**, która będzie przedstawiała powierzchnię danego kraju w danym roku w milionach km^2. Oprócz tego dodajmy także mapowanie kolorów odpowiednie dla daltonistów.

```r
data$GDP <- data$GDPPPP_percapita * data$Population * 10^6
data$Area <- data$Population / data$Population_density
colors <- c('Kumulatywne' = 'red',
            '1960' = rgb(180, 40, 50, maxColorValue = 255),
```

```
            '1980' = rgb(140, 100, 100, maxColorValue = 255),
            '2000' = rgb(100, 160, 150, maxColorValue = 255),
            '2015' = rgb(60, 220, 200, maxColorValue = 255))
data_1960 <- data %>%
  filter(Year == 1960)
data_1980 <- data %>%
  filter(Year == 1980)
data_2000 <- data %>%
  filter(Year == 2000)
data_2015 <- data %>%
  filter(Year == 2015)
```

1. Współzależność Urbanizacji i Śmiertelności Dziecięcej

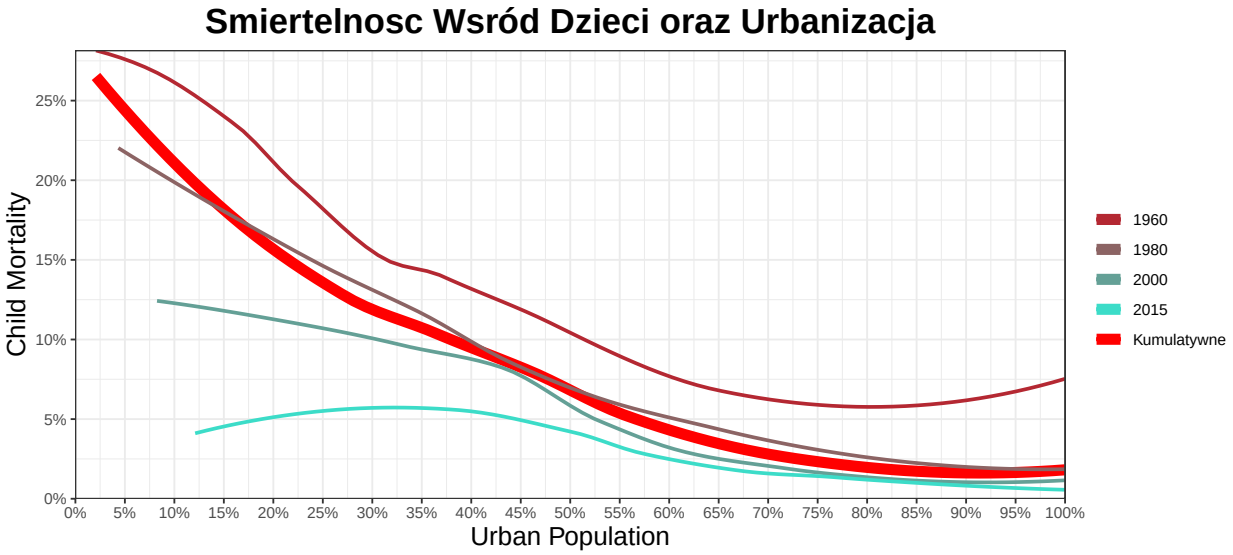
Ciekawą kwestią jest **śmiertelność dziecięca** w krajach o różnych stopniach **urbanizacji**. Sekcja ta będzie się zajmować badaniem tej zależności.

Przejdźmy najpierw do sprawdzenia jak wygląda trend dla kilku wybranych lat.

```
ggplot() +
  geom_smooth(aes(Urban_population / 100, Child_mortality / 100,
                  color = 'Kumulatywne'),
              data, se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(Urban_population / 100, Child_mortality / 100,
                  color = '1960'), data_1960, se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Child_mortality / 100,
                  color = '1980'), data_1980, se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Child_mortality / 100,
                  color = '2000'), data_2000, se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Child_mortality / 100,
                  color = '2015'), data_2015, se = F,
              method = 'loess', formula = y ~ x) +
  ggtitle('Śmiertelność Wsród Dzieci oraz Urbanizacja') +
  labs(x = 'Urban Population',
       y = 'Child Mortality',
       color = NULL) +
  expand_limits(x = 0, y = 0) +
  scale_x_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  scale_y_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15)) +
  scale_color_manual(values = colors)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
```

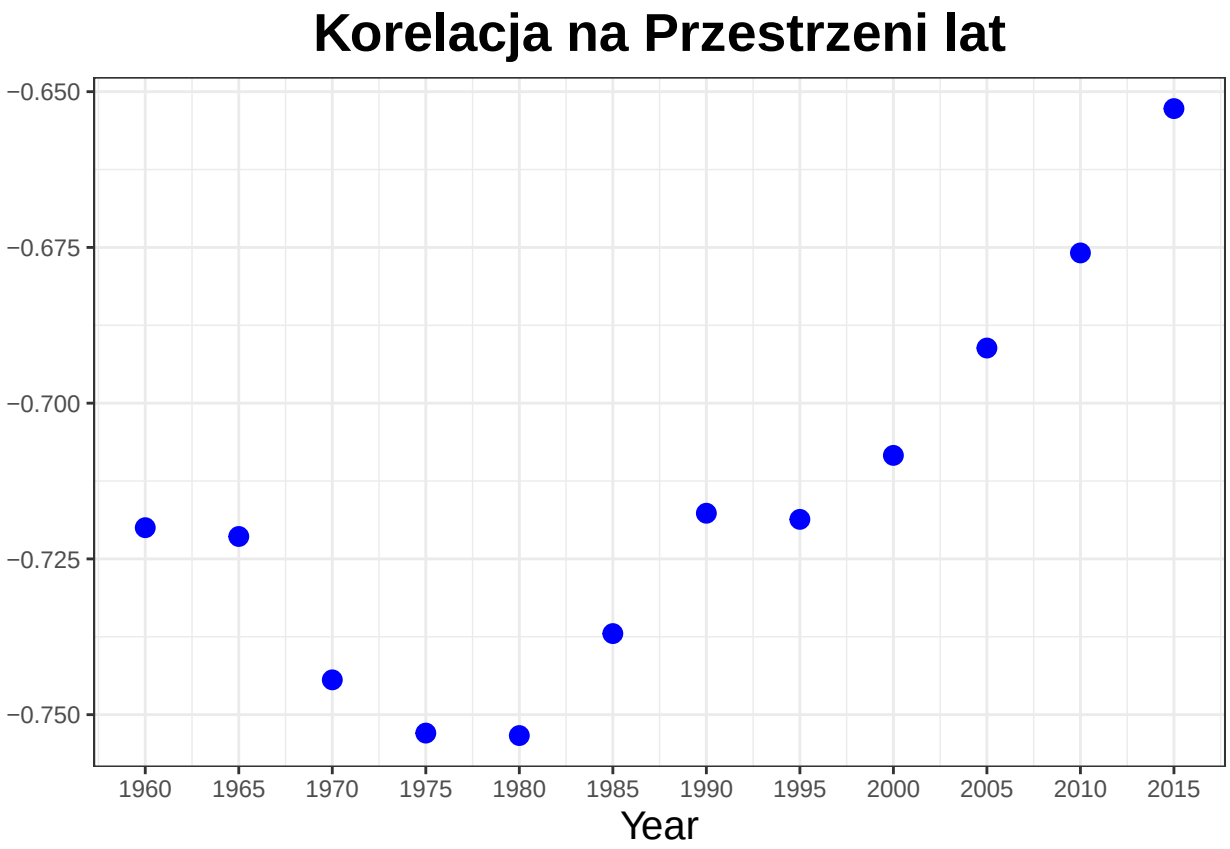
```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



W prawie wszystkich wybranych latach ogólny trend jest ściśle ujemny, jednak w roku 2015 widzimy odcinek na którym trend jest dodatni. Możliwe jest, że istnieje pewna zależność pomiędzy rokiem z którego pochodzą dane, a współzależnością tych dwóch zmiennych.

Przejdźmy teraz do sprawdzenia tego przypuszczenia.

```
data %>%
  group_by(Year) %>%
  summarise(Korelacja = cor(Urban_population, Child_mortality,
                            method = 'spearman')) %>%
ggplot() +
  geom_point(aes(Year, Korelacja), color = 'blue', size = 3) +
  ggtitle('Korelacja na Przestrzeni lat') +
  labs(y = NULL) +
  scale_x_continuous(breaks = seq(1960, 2015, 5)) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15))
```



Widzimy tutaj ewidentny ciąg rosnący począwszy od roku 1980.

Podsumowanie

Chociaż nie do zaprzeczenia jest pewna odwrotna współzależność między **śmiertelnością dziecięcą**, a **urbanizacją** to widzimy, że siła tej zależności od kilku dziesięciu lat znacząco maleje. Wygląda na to, że mieszkanie na terenie wiejskim nie wróży już takiego strasznego przeznaczenia dla dzieci jak kiedyś. Postawiona hipoteza nie została spełniona w całości, gdyż zależność pomiędzy tymi dwoma zmiennymi jest silna lub umiarkowanie silna wzależności od roku badania, jednak kierunek tej zależności jest generalnie jednostronny.

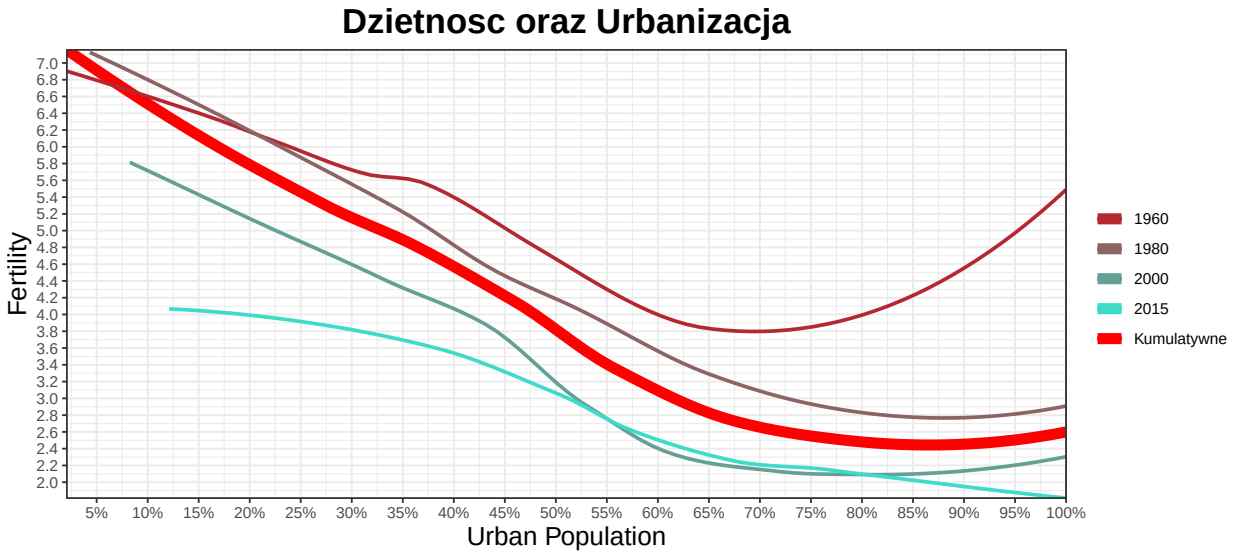
2. Zależność Między Dzietnością a Urbanizacją

Wiemy już, że **urbanizacja** oraz **śmiertelność dziecięca** są przeciwnie skorelowane, chociaż siła tej zależności maleje z biegiem lat. Przejdźmy teraz do sprawdzenia czy ta sama zależność zachodzi pomiędzy **dzietnością**, a **urbanizacją**. **Dzietność** oraz **śmiertelność dziecięca** mają wysoki współczynnik korelacji, więc można się spodziewać podobnych wyników, jednakże przystąpmy do dokładnej analizy.

Spróbujmy najpierw wyodrębnić trend.

```
ggplot() +
  geom_smooth(aes(Urban_population / 100, Fertility,
                  color = 'Kumulatywne'),
              data, se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(Urban_population / 100, Fertility,
                  color = '1960'), data_1960, se = F,
```

```
            method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Fertility,
                  color = '1980'), data_1980, se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Fertility,
                  color = '2000'), data_2000, se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Urban_population / 100, Fertility,
                  color = '2015'), data_2015, se = F,
              method = 'loess', formula = y ~ x) +
  ggtitle('Dzietność oraz Urbanizacja') +
  labs(x = 'Urban Population',
       y = 'Fertility',
       color = NULL) +
  scale_x_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  scale_y_continuous(expand = c(0, NA), breaks = seq(0, 10, 0.2)) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15)) +
  scale_color_manual(values = colors)
```



Widzimy tutaj sytuację nieco odwrotną w porównaniu do poprzedniego tematu. Trend nie jest ściśle malejący dla danych najstarszych. Spróbujmy teraz zobrazować korelację na przestrzeni lat.

```
corr_summary <- data %>%
  group_by(Year) %>%
  summarise(Korelacja = cor(Urban_population, Fertility, method = 'spearman'))

ggplot(corr_summary) +
  geom_point(aes(Year, Korelacja), color = 'blue', size = 3) +
  ggtitle('Korelacja na Przestrzeni lat') +
  labs(y = NULL) +
  scale_x_continuous(breaks = seq(1960, 2015, 5)) +
  theme_bw() +
```

```
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15))
```

Mimo tego, że trend wygląda nie co inaczej korelacja wynikowa wyszła podobna. Przynajmniej można tak powiedzieć o latach 1985-2015. Widzimy jednak, że korelacja bezwzględna była znacząco niższa w latach wcześniejszych, co nie miało miejsca w poprzednio rozważanej sytuacji. Można więc wnioskować, że istnieje czynnik, który ma silniejszą współzależność z **dzietnością** niż ze **śmiertelnością dzieciącą**. Może to być spowodane faktem, że osoby w skrajnie złej sytuacji majątkowej bardziej powszechnie znajdywały się na teranach miejskich, jest to jednak tylko przypuszczenie.

Podsumowanie

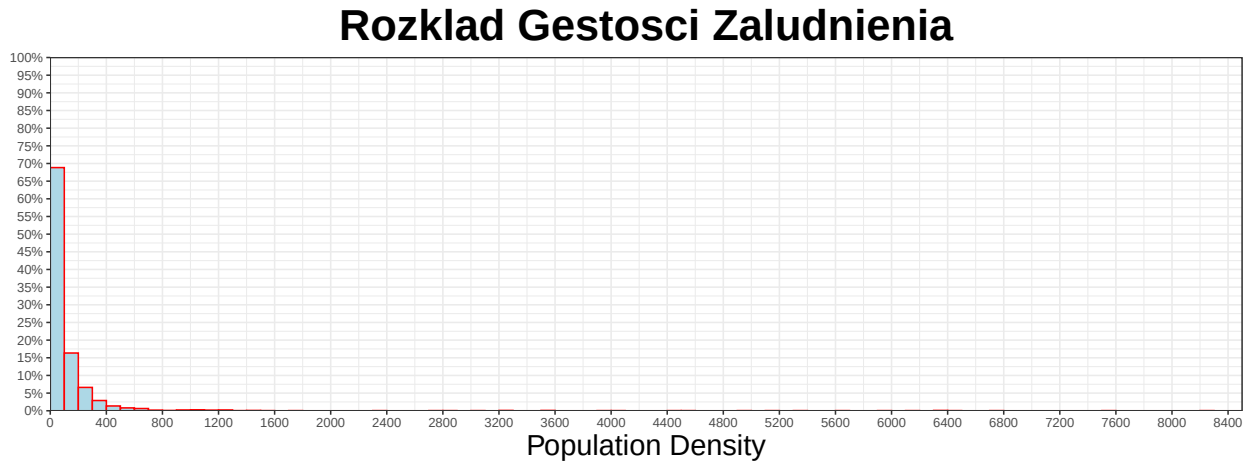
Widzimy więc, że sytuacje w przypadku **śmiertelności dziecięcej** oraz **dzietności** są podobne. Główną różnicą pomiędzy nimi jest fakt, że korelacja zachowywała się inaczej przez rokiem 1985.

3. Relacja Pomiędzy Gęstością Zaludnienia a Nierównościami

Kontynuujemy więc nasze rozważania dotyczące relacji między rozłożeniem ludności, a statystykami ekonomiczymi/społecznymi. Tym razem obiektem badań będzie charakter relacji pomiędzy **gęstością zaludnienia**, a nierównościami społecznymi (zobrazowanymi za pomocą **współczynnika Giniego**). Z macierzy korelacji widzimy, że sama korelacja w całym zestawie danych jest tutaj dużo niższa niż w przypadku poprzedniej pary zmiennych. Nie dochodźmy jednak do zbyt pochobnych wniosków. Przeprowadźmy analizę trendu.

Zbadajmy najpierw rozkład realnego dochodu.

```r
ggplot(data) +
  geom_histogram(aes(Population_density,
                     after_stat(count)/sum(after_stat(count))),
                 boundary = 0,
                 binwidth = 100,
                 fill = 'light blue', color = 'red') +
  ggtitle('Rozkład Gęstości Zaludnienia') +
  expand_limits(x = c(0, max(data$Population_density) + 200), y = c(0, 1)) +
  scale_y_continuous(breaks = seq(0, 1, 0.05), expand = c(0, NA),
                     labels = scales::percent) +
  scale_x_continuous(expand = c(0, NA), breaks = seq(0, 10000, 400)) +
  labs(x = 'Population Density',
       y = NULL) +
  theme_bw() +
  theme(plot.title = element_text(size = 30, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 20))
```

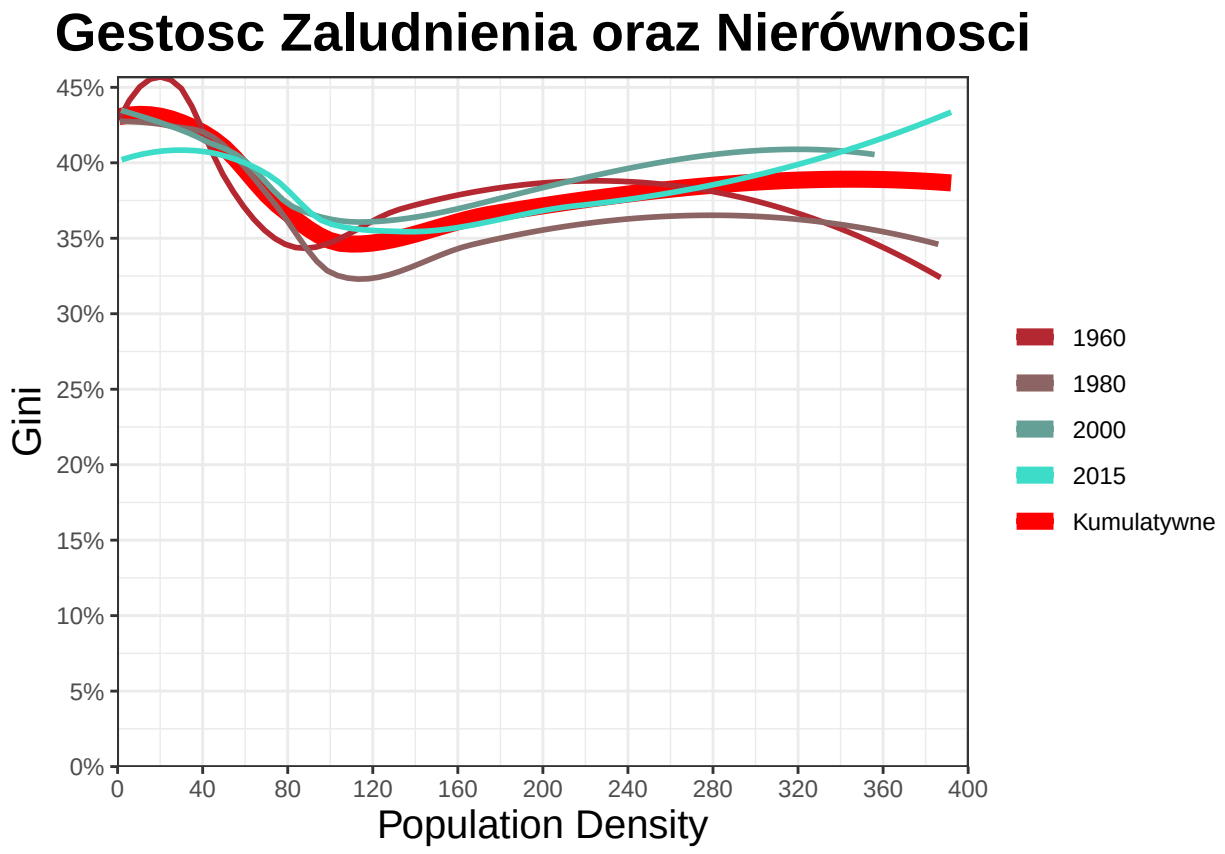


Na wykresie widzimy, że **gęstość populacji** jest zmienną wysoko skoncentrowaną, tzn. większość obserwacji znajduję się w przedziale małym stosunkowo do odchylenia standardowego. Rozsądne będzie więc zawężenie przedziału w którym będziemy przeprowadzać analizę do [0, 400].

Spróbujmy teraz wyodrębnić trend w nim.

```r
ggplot() +
  geom_smooth(aes(Population_density, Gini / 100,
                  color = 'Kumulatywne'),
              data %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(Population_density, Gini / 100,
                  color = '1960'), data_1960 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Gini / 100,
                  color = '1980'), data_1980 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Gini / 100,
                  color = '2000'), data_2000 %>%
```

```
              filter(Population_density <= 400), se = F,
            method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Gini / 100,
                  color = '2015'), data_2015 %>%
              filter(Population_density <= 400), se = F,
            method = 'loess', formula = y ~ x) +
  ggtitle('Gęstość Zaludnienia oraz Nierówności') +
  labs(x = 'Population Density',
       y = 'Gini',
       color = NULL) +
  expand_limits(x = c(0,400), y = 0) +
  scale_x_continuous(expand = c(0, NA),breaks = seq(0, 400, 40)) +
  scale_y_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15)) +
  scale_color_manual(values = colors)
```



Widzimy, że generalny trend jest malejący w przedziale [0, 100] oraz rosnący w przedziale [100, 400]. Sprawdźmy teraz względną liczebność obu tych przedziałów.

```
cat('Przedział [0,100]:', sum(data$Population_density <= 100) /
      length(data$Population_density), '\n')
cat('Przedział [100,400]:', sum(data$Population_density >= 100 &
```

```
                                data$Population_density <= 400) /
      length(data$Population_density), '\n')
```

```
## Przedział [0,100]: 0.6884058
## Przedział [100,400]: 0.259058
```

Widzimy, że większość obserwacji zawiera się w przedziale [0, 100]. Duża ich część znajduje się jednak, także w przedziale [100, 400].

Podsumowanie

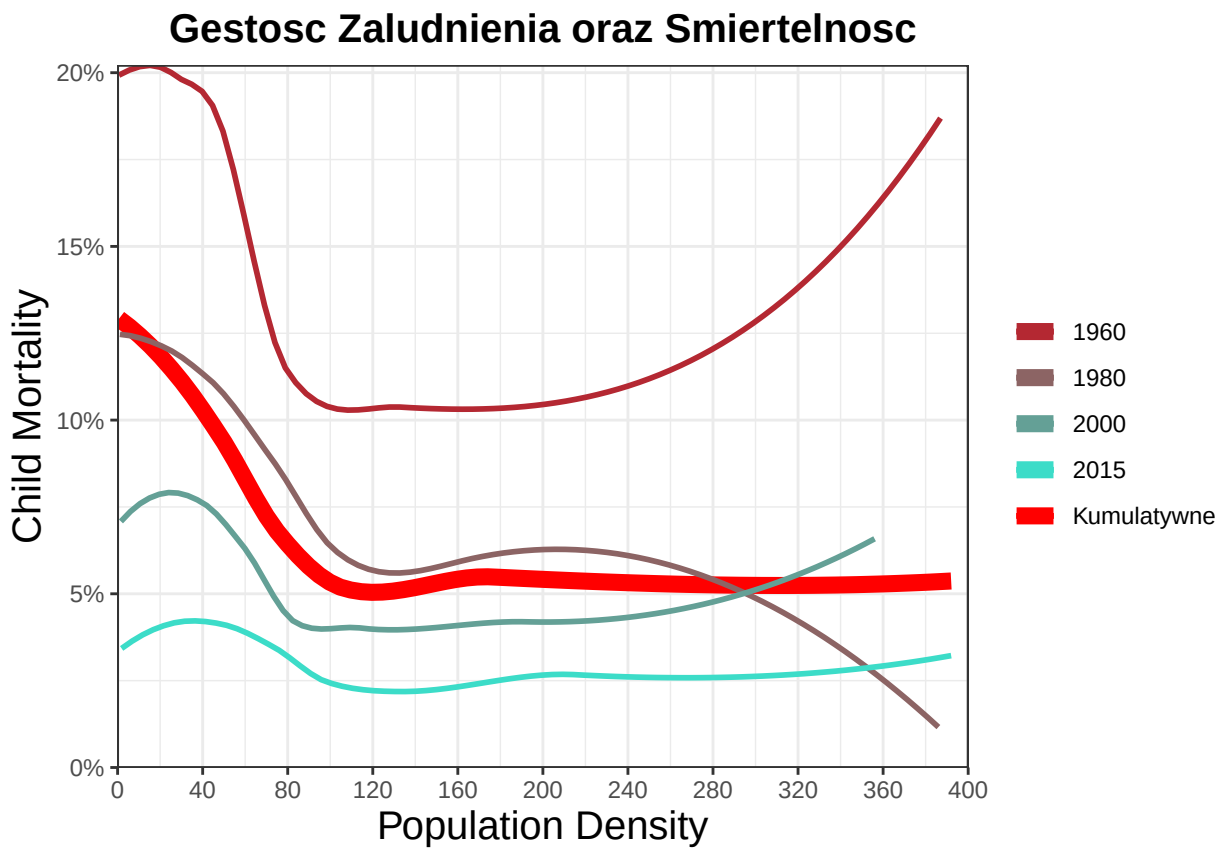
Z powyższych rozważań można wnioskować, że na ujemną wynikową korelację składa się przede wszystkim znaczący ujemny trend w przedziale [0, 100]. Nie jest to jednak koniec historii, gdyż trend jest generalnie lekko rosnący w [100, 400]. Ostatecznie kwestia zależności między **gęstością zaludnienia**, a **współczynnikiem Giniego** jest nietrywialna. Hipoteza jest generalnie tylko spełniona dla przedziału [100, 400].

4. Gęstość Zaludnienia a Śmiertelność Dzieci

Gęstość zaludnienia może jednak być mocniej skorelowana z innymi zmiennymi. Macierz korelacji wskazuje na to, że **śmiertelność dzieci** jest tu najbardziej z nią współzależną (na równo z **lifeExp**). Możemy tutaj więc oczekiwać silniejszej zależności niż w poprzednim przypadku. Spróbujmy wyodrębnić trend.

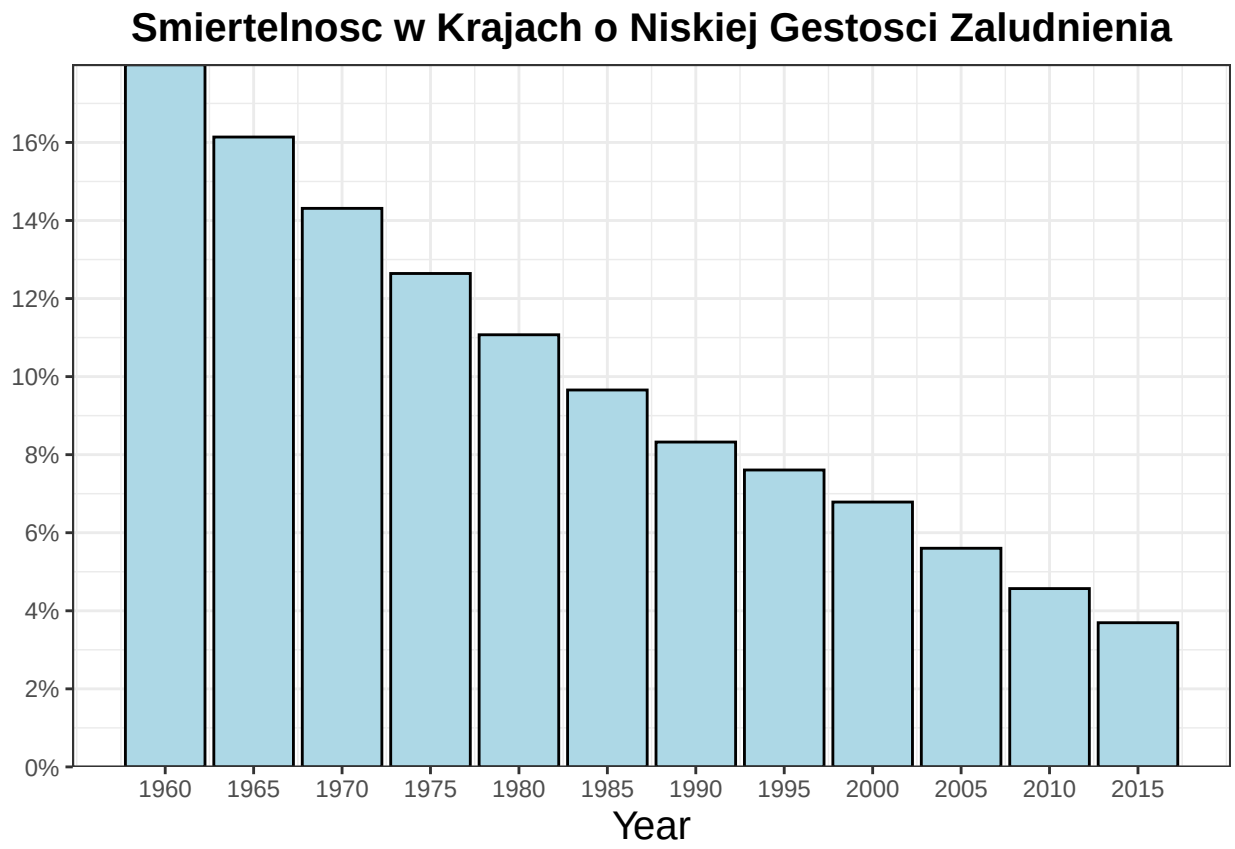
```
ggplot() +
  geom_smooth(aes(Population_density, Child_mortality / 100,
                  color = 'Kumulatywne'),
              data %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(Population_density, Child_mortality / 100,
                  color = '1960'), data_1960 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Child_mortality / 100,
                  color = '1980'), data_1980 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Child_mortality / 100,
                  color = '2000'), data_2000 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Population_density, Child_mortality / 100,
                  color = '2015'), data_2015 %>%
                filter(Population_density <= 400), se = F,
              method = 'loess', formula = y ~ x) +
  ggtitle('Gęstość Zaludnienia oraz Śmiertelność') +
  labs(x = 'Population Density',
       y = 'Child Mortality',
       color = NULL) +
  expand_limits(x = c(0, 400), y = 0) +
  scale_x_continuous(expand = c(0, NA),breaks = seq(0, 400, 40)) +
  scale_y_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  theme_bw() +
  theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
```

```
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15)) +
  scale_color_manual(values = colors)
```



Widzimy tutaj podobne zjawisko jak poprzednio, trend jest mocno malejący w przedziale [0, 100]. Nie widać jednak widocznego trendu w przedziale [100, 400] w przeciwieństwie do **współczynnika Giniego**. Zbadajmy jeszcze dodatkowo jak zmieniła się średnia **śmiertelność dziecięca** w krajach nisko zaludnionych (≤ 100 os. na km^2) na przestrzeni lat.

```
data %>%
  filter(Population_density <= 100) %>%
  ggplot() +
    stat_summary(aes(Year, Child_mortality / 100), fun = mean, geom = 'bar',
                 fill = 'light blue', color = 'black') +
    ggtitle('Śmiertelność w Krajach o Niskiej Gęstości Zaludnienia') +
    labs(y = NULL) +
    scale_y_continuous(breaks = seq(0, 1, 0.02), expand = c(0, NA),
                       labels = scales::percent) +
    scale_x_continuous(breaks = seq(1960, 2015, 5)) +
    theme_bw() +
    theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
          axis.title.x = element_text(size = 15))
```



Widzimy w tym szeregu czasowym trend znacząco malejący. Dodatkowo możemy zauważyć, że **Śmiertelność dziecięca** spadła poniżej 4% w roku 2015.

Podsumowanie

Widzimy, że w krajach o bardzo niskiej **gęstości** zaludnienia śmiertelność maleje wraz ze wzrostem tej zmiennej. Ciężko jest jednak mówić o trendzie w pozostałej części danych. Przedział [100, 400] wykazuje brak zależności, a pozostałe dane są zbyt nieliczne aby wyciągać wnioski. Możemy więc wnioskować, że hipoteza jest generalnie fałszywa. Dodatkowo na podstawie ostatniego wykresu możemy dojść do wniosku, że **śmiertelność dziecięca** zmalała znacząco w najrzadziej zaludnionych krajach przez badany okres, jest to ewidentnie skutek rozwoju cywilizacyjnego.

5. Zależność Między Nierównościami Społecznymi i Dochodem

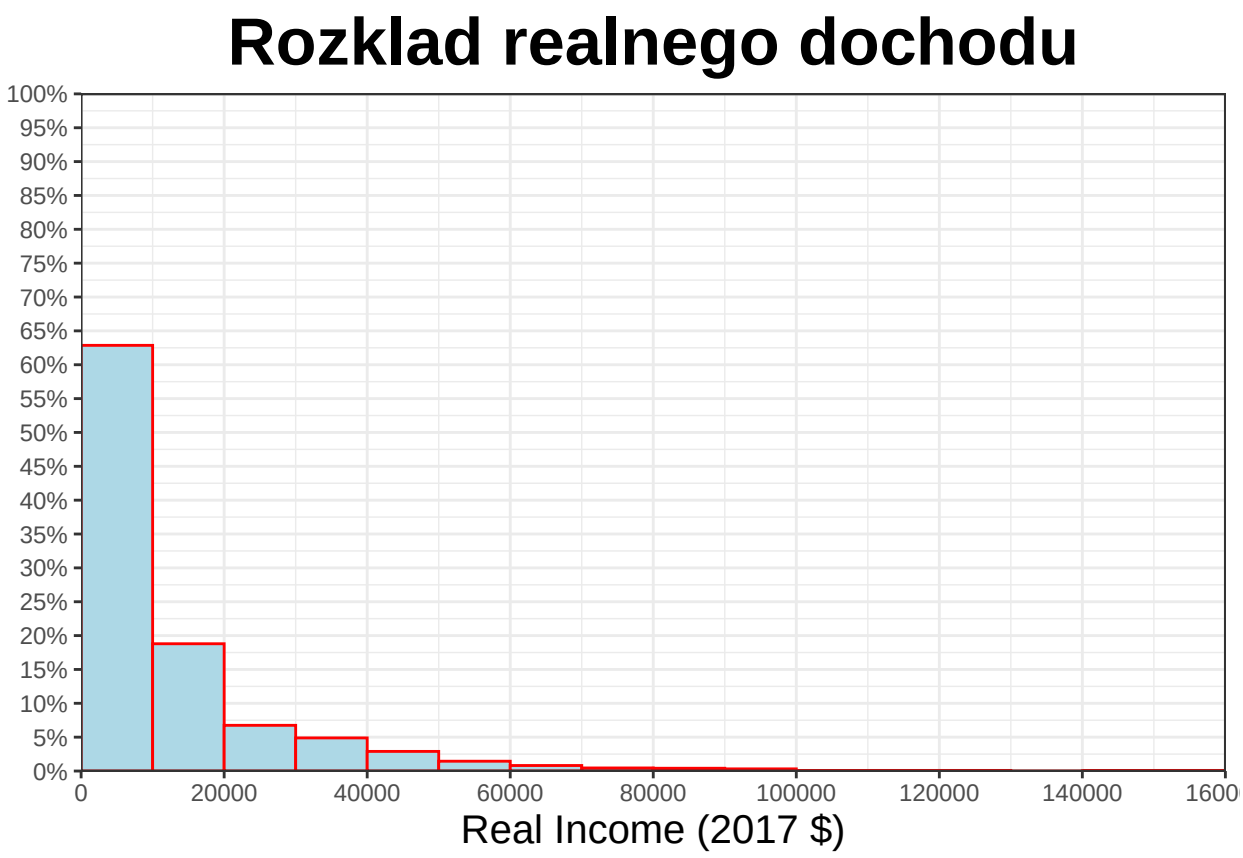
Wróćmy teraz do tematu nierówności dochodowych. Macierz korelacji od razu sugeruje, że współzależność między **współczynnikiem Giniego**, a **realnym dochodem** (jeżeli istnieje) jest nikła. Powinniśmy przed dokonaniem zbyt wczesnego wnioskowania, przyjrzeć się tymi danymi dokładniej.

W tym celu stwórzmy histogram obrazujący rozkład **Realnego Dochodu**, gdyż nią będziemy traktować jako zmienną niezależną.

```
ggplot(data) +
  geom_histogram(aes(GDPPPP_percapita, ..count../sum(..count..)),
                 boundary = 0,
                 binwidth = 10000,
                 fill = 'light blue', color = 'red') +
  ggtitle('Rozkład realnego dochodu') +
```

```
  expand_limits(y = c(0, 1)) +
  scale_y_continuous(breaks = seq(0, 1, 0.05), expand = c(0, NA),
                     labels = scales::percent) +
  scale_x_continuous(expand = c(0, NA), breaks = seq(0, 160000, 20000)) +
  labs(x = 'Real Income (2017 $)',
       y = NULL) +
  theme_bw() +
  theme(plot.title = element_text(size = 25, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15))
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

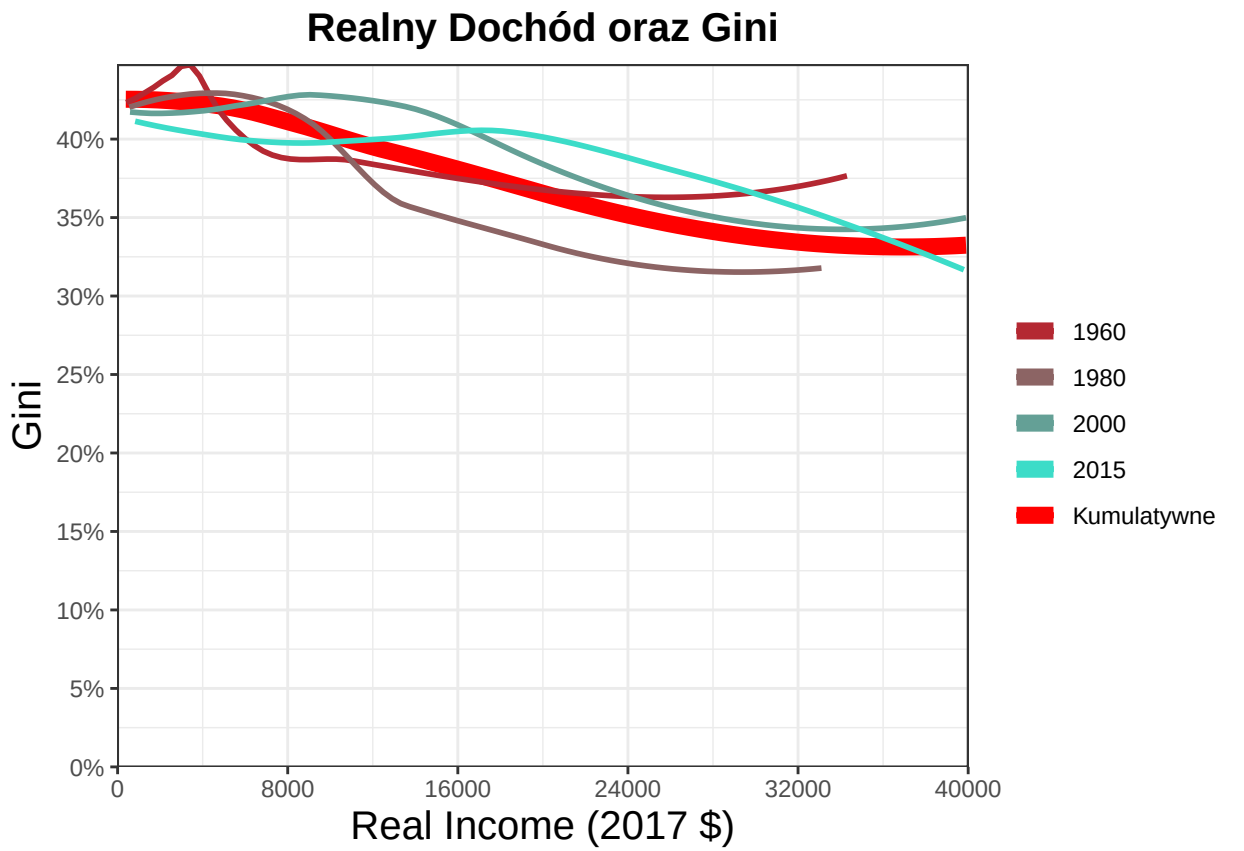


Widzimy więc, że będzie rozsądne wybranie przedzialu [0, 40000] w celu otrzymania reprezentatywnej próby.

Spróbujmy teraz wyodrębnić trend dla tego przedziału.

```
ggplot() +
  geom_smooth(aes(GDPPPP_percapita, Gini / 100,
                  color = 'Kumulatywne'),
              data %>%
                filter(GDPPPP_percapita <= 40000), se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(GDPPPP_percapita, Gini / 100,
```

```
                color = '1960'), data_1960 %>%
              filter(GDPPPP_percapita <= 40000), se = F,
            method = 'loess', formula = y ~ x) +
geom_smooth(aes(GDPPPP_percapita, Gini / 100,
                color = '1980'), data_1980 %>%
              filter(GDPPPP_percapita <= 40000), se = F,
            method = 'loess', formula = y ~ x) +
geom_smooth(aes(GDPPPP_percapita, Gini / 100,
                color = '2000'), data_2000 %>%
              filter(GDPPPP_percapita <= 40000), se = F,
            method = 'loess', formula = y ~ x) +
geom_smooth(aes(GDPPPP_percapita, Gini / 100,
                color = '2015'), data_2015 %>%
              filter(GDPPPP_percapita <= 40000), se = F,
            method = 'loess', formula = y ~ x) +
ggtitle('Realny Dochód oraz Gini') +
labs(x = 'Real Income (2017 $)',
     y = 'Gini',
     color = NULL) +
expand_limits(x = c(0, 40000), y = 0) +
scale_x_continuous(expand = c(0, NA), breaks = seq(0, 40000, 8000)) +
scale_y_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.05),
                   labels = scales::percent) +
theme_bw() +
theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
      axis.title.x = element_text(size = 15),
      axis.title.y = element_text(size = 15)) +
scale_color_manual(values = colors)
```



Generalnie widzimy tutaj, że w przypadku większości krajów Większy **realny dochód** zazwyczaj implikuje mniejszą wartość **współczynnika Giniego**. Trend jest jednak tylko lekko malejący co może rodzić co do jego istotności.

Podsumowanie

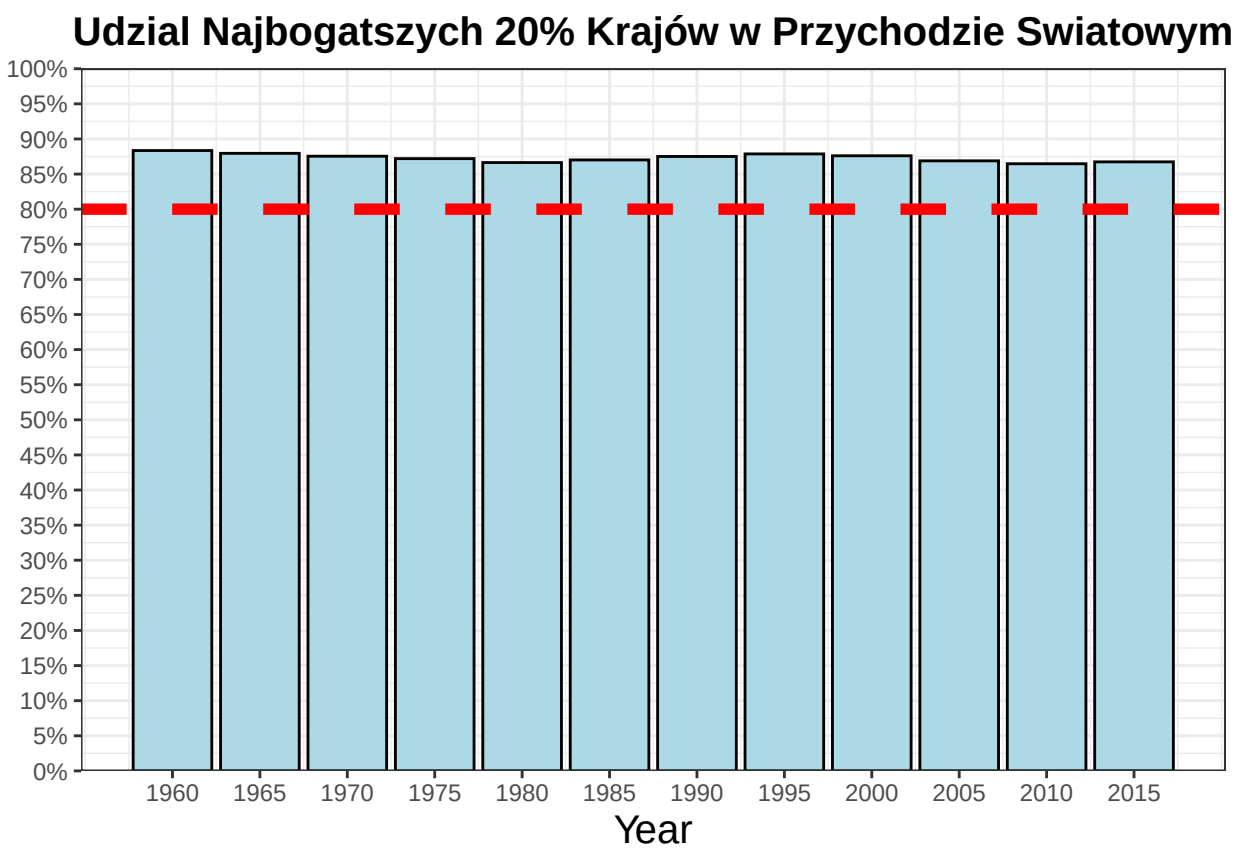
Po głębszej analizie możemy dojść do wniosku, że mimo, że pewna korelacja pomiędzy tymi dwoma zmiennymi istnieje w przypadku większości społeczeństw to nie powinniśmy traktować tego jako prawdę absolutną. Współzależność, którą znaleźliśmy jest nikła jak na to już wskazywał korelogram przedstawiony na początku tej pracy. Oczywiście nie znaczy to, że wnioski, które otrzymaliśmy są bezużyteczne. Współzależność tych dwóch czynników nadal jest czymś co powinniśmy wziąść pod uwagę w badaniu zjawisk ekonomicznych.

6. Udział Najbogatszych Krajów w Dobytku

Poprzednio badaliśmy nierówności wewnątrz krajów, zbadajmy teraz nierówności między krajami. Statystyką badaną będzie tutaj **kumulatywny przychód mieszkańców kraju według parytetu siły nabywczej**. Stwórzymy najpierw wykres kolumnowy obrazujący udział najbogatszych 20% krajów w produkcie światowym na przestrzeni lat.

```
ggplot(data) +
 stat_summary(aes(Year, GDP), fun = function(x)
   {return(sum(x[ x > quantile(x, probs = seq(0, 1, 0.1))[9]])/sum(x))},
   geom = 'bar', fill = 'light blue',
              color = 'black') +
  ggtitle('Udział Najbogatszych 20% Krajów w Przychodzie Światowym') +
  labs(y = NULL) +
```

```
  expand_limits(y = c(0, 1)) +
  scale_y_continuous(breaks = seq(0, 1, 0.05), labels = scales::percent,
                     expand = c(0, NA)) +
  scale_x_continuous(breaks = seq(1960, 2015, 5)) +
  geom_hline(yintercept = 0.8, linetype = 'dashed', color = 'red', size = 2) +
  theme_bw() +
  theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15))
```



Jak widzimy we wszystkich badanych latach hipoteza jest spełniona, dodatkowo w podobnym stopniu. W większości lat najbogatsze 20% krajów nie posiada jedynie 80% dobytku światowego, ale ponad 85%. Jednak pozostaje nam jeszcze zbadanie czy to bogactwo przypada na tą samą część ludności światowej. Zanim jednak do tego przystąpimy to powinniśmy sprawdzić ile krajów w najbogatszych 20% się powtarza na przestrzeni lat. Dobrze by było także wiedzieć jakie to są kraje.

```
same_countries <- data_1960$Country[data_1960$GDP >
                                    quantile(data_1960$GDP,
                                             probs = seq(0, 1, 0.1))[9]]
for (i in seq(1965, 2015, 5)) {
  same_countries <- intersect(same_countries, (filter(data, Year == i) %>%
                              filter(GDP >
                                       quantile(GDP,
                                        probs = seq(0, 1, 0.1))[9]))$Country)
}

cat('Proporcja tych samych krajów:', length(same_countries) /
```

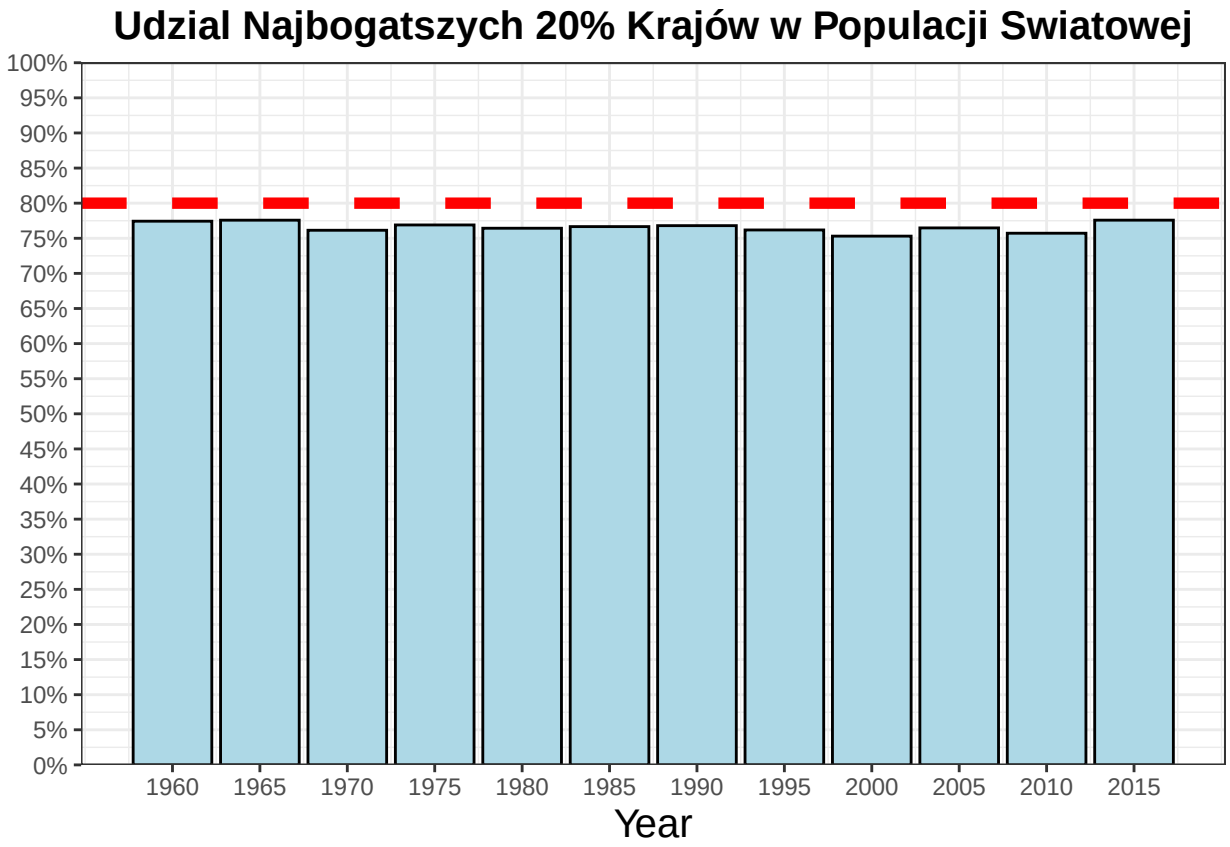
```
      floor(0.2 * length(data_1960$Country)), '\n')
cat('Powtarzające się kraje:\n')
same_countries
rm(i)
```

```
## Proporcja tych samych krajów: 0.7222222
## Powtarzające się kraje:
##  [1] "Argentina"      "Australia"      "Belgium"        "Brazil"
##  [5] "Canada"         "China"          "Colombia"       "France"
##  [9] "Germany"        "India"          "Indonesia"      "Iran"
## [13] "Italy"          "Japan"          "Mexico"         "Netherlands"
## [17] "Poland"         "Russia"         "South Africa"   "Spain"
## [21] "Switzerland"    "Turkey"         "Ukraine"        "United Kingdom"
## [25] "United States"  "Venezuela"
```

Widzimy, że ponad $\frac{2}{3}$ krajów się powtarza w najbogatszych 20%. Wśród tych krajów widzimy głównie kraje środkowo-europejskie oraz kilka krajów bliskiego wschodu. Pojawiły się tu jednak też kraje z Ameryki Łacińskiej, a także Chiny oraz Indie. Fakt, że te dwa ostatnie się tu znalazły jest w największym stopniu spowodowane ich wysoką populacją. Kraje pierwszego świata znajdują się na tej liście z powodów raczej oczywistych. Kraje bliskiego wschodu znalazły się tu natomiast z powodu ich oligopolu na rynku ropy naftowej.

Przejdźmy teraz do zobrazowania jaka cześć populacji mieszka w najbogatszych krajach.

```
data %>%
  arrange(desc(GDP)) %>%
  ggplot() +
    stat_summary(aes(Year, Population), fun = function(x)
     {return(sum(x[1:floor(length(x) * 0.2)])/sum(x))},
     geom = 'bar',
                fill = 'light blue', color = 'black') +
    ggtitle('Udział Najbogatszych 20% Krajów w Populacji Światowej') +
    labs(y = NULL) +
    expand_limits(y = c(0, 1)) +
    scale_y_continuous(breaks = seq(0, 1, 0.05), labels = scales::percent,
                       expand = c(0, NA)) +
    scale_x_continuous(breaks = seq(1960, 2015, 5)) +
    geom_hline(yintercept = 0.8, linetype = 'dashed', color = 'red', size = 2) +
    theme_bw() +
    theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
          axis.title.x = element_text(size = 15))
```



Widzimy więc, że w każdym roku badania na 20% najbogatszych krajów przypada na mniej niż 80% populacji.

Podsumowanie

Ostatecznie więc wiemy:

- 20% najbogatszych krajów posiada ponad 80% (a nawet prawie 90%) dobytku światowego, co potwierdza naszą hipotezę. Dodatkowo na pierwszym przedstawionym wykresie widzimy, że ten odsetek nie zmienia się zbytnio na przestrzeni lat;
- Ponad $\frac{2}{3}$ badanych krajów się powtarza we wszystkich badanych latach. Zazwyczaj w czołówce znajdują się więc te same kraje. Oznacza to, że kraje biedniejsze nie mają zbytnio szansy na zmianę swojej sytuacji;
- Na 20% najbogatszych krajów przypada jednak mniej niż 80% ludności. Bezpośrednio płynącym z tego wnioskiem jest fakt, że na ponad 20% populacji przypada jedynie około kilkunastu procent kumulatywnej siły nabywczej.

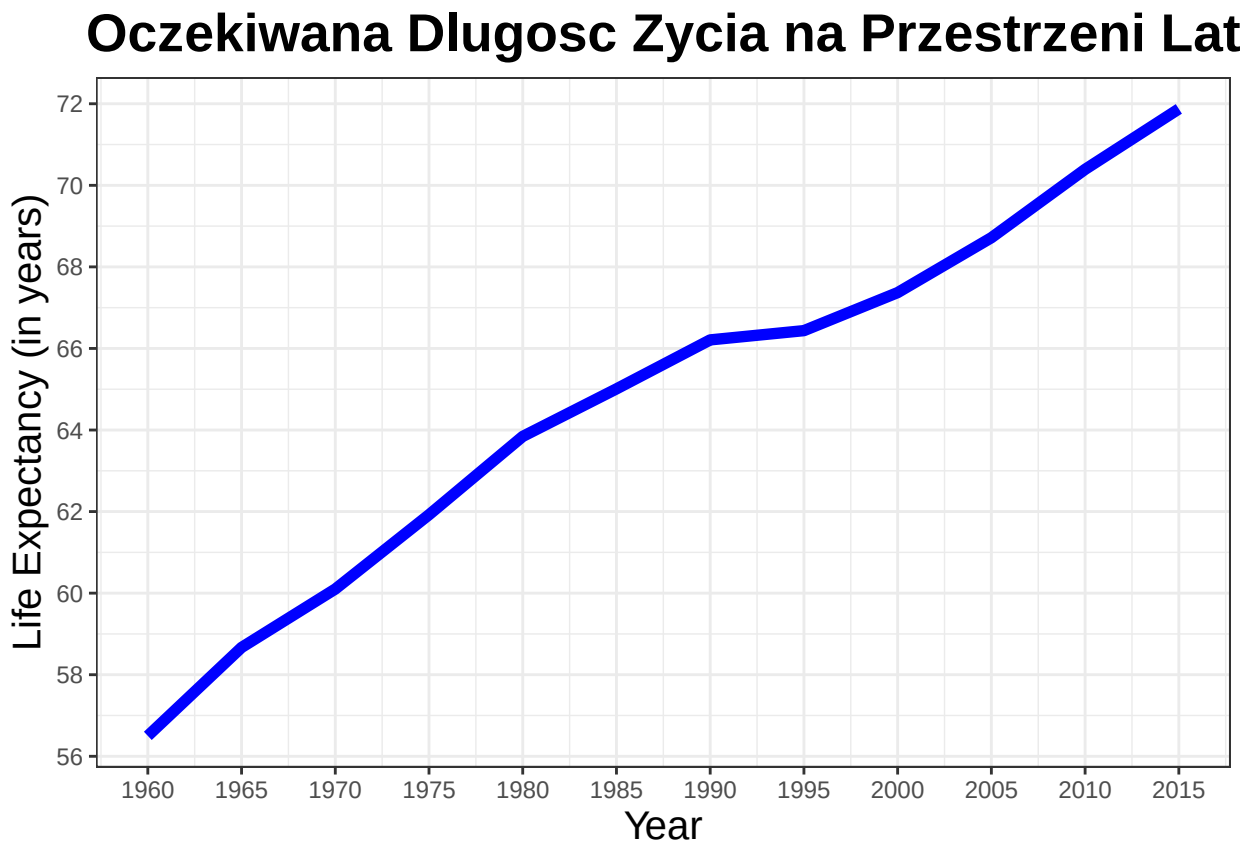
7. Oczekiwana Długość życia a Dzietność

Odejdźmy już od danych typowo ekonomiczno-geograficznych, a skupmy się na badaniu zmiennych losowych dotyczących życia mieszkańców danego kraju.

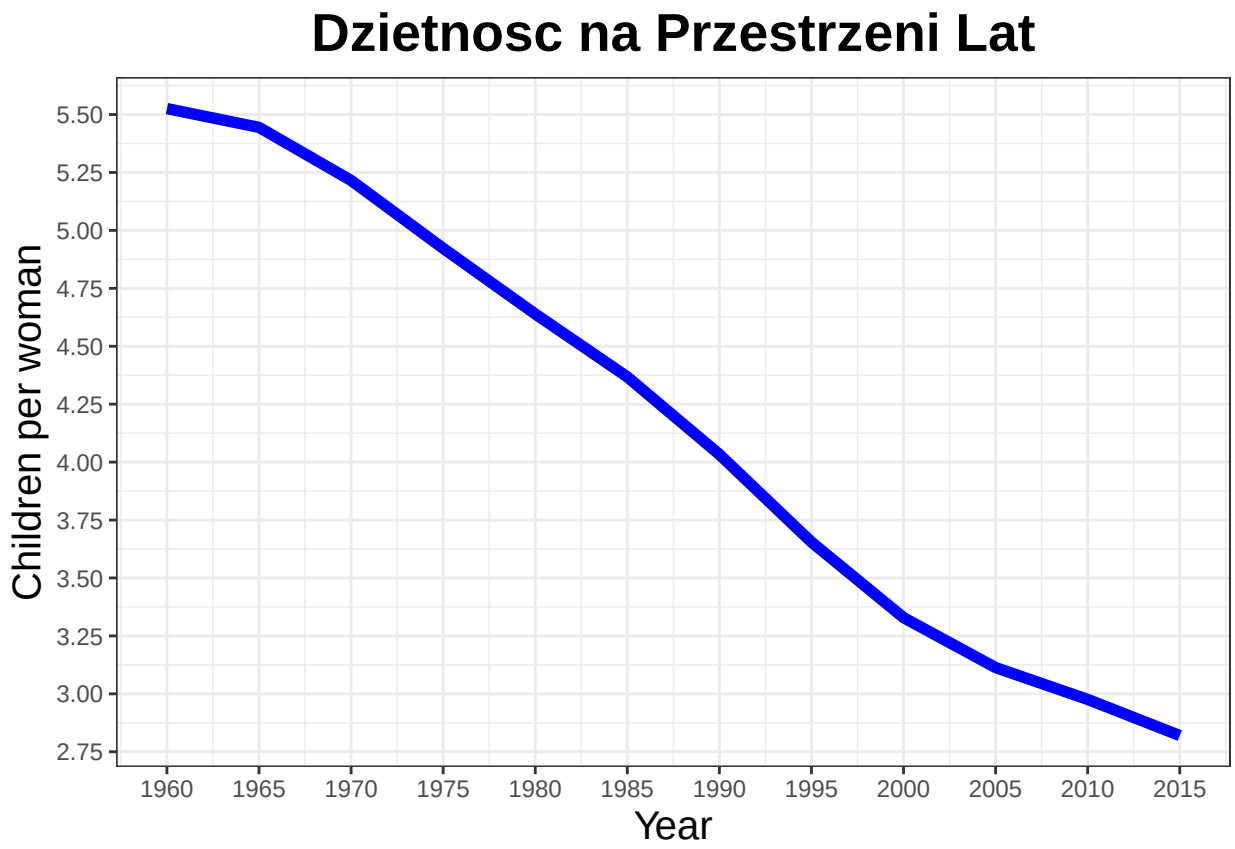
Kwestią niekontrowersyjną jest raczej, że średnia **długość życia** rośnie wraz z rozwojem cywilizacyjnym, spodziewamy się także, że **dzietność** będzie malała.

Zobrazujemy teraz średnie wartości obydwu tych zmiennych na przestrzeni lat.

```
ggplot(data) +
  geom_line(aes(Year, Life_exp),
            stat = 'summary', fun = mean, color = 'blue', size = 2) +
  ggtitle('Oczekiwana Długość Życia na Przestrzeni Lat') +
  labs(y = 'Life Expectancy (in years)') +
  scale_x_continuous(breaks = seq(1960, 2015, 5)) +
  scale_y_continuous(breaks = seq(50, 80, 2)) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15))
```




```
ggplot(data) +
  geom_line(aes(Year, Fertility),
            stat = 'summary', fun = mean, color = 'blue', size = 2) +
  ggtitle('Dzietność na Przestrzeni Lat') +
  labs(y = 'Children per woman') +
  scale_x_continuous(breaks = seq(1960, 2015, 5)) +
  scale_y_continuous(breaks = seq(0, 10, 0.25)) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15))
```



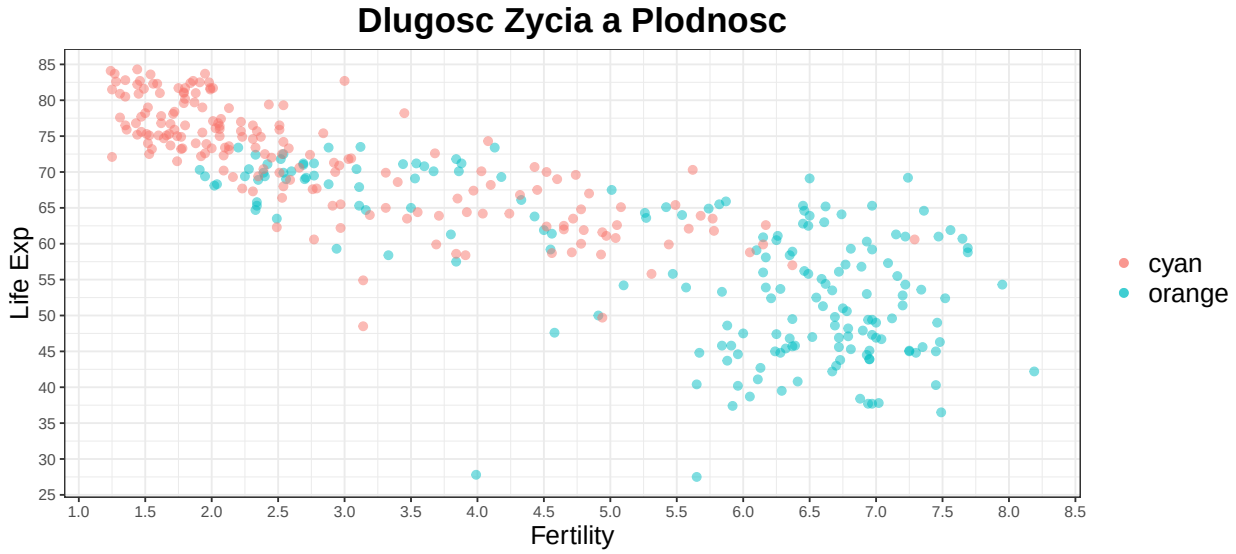
Widzimy, że nasze przypuszczenia co do trendu tych dwóch zmiennych względem czasu są prawdziwe. Możemy odczytać także, że **oczekiwana długość życia** wzrosła o 15 lat, natomiast **płodność** zmalała 2-krotnie. Ostatecznie widzimy, że obydwa wykresy przedstawiają zależność w przybliżeniu liniową. Nie może to być zależność liniowa, gdyż **oczekiwana długość życia** nie może rosnąć w nieskończoność (a przynajmniej nie z trendem liniowym). Natomiast **Płodność** nie może mieć trendu liniowego ze względu na fakt że zmienna ta jest ograniczona od dołu (musi być dodatnia).

Zobrazujmy teraz korelację tych dwóch zmiennych za pomocą wykresu punktowego.

Wyodrębnijmy tutaj jednak tylko dwa skrajne lata 1960 oraz 2015 aby nie utrudniać interpretacji wykresu.

```
ggplot() +
  geom_point(aes(Fertility, Life_exp, color = 'orange'), alpha = 0.5,
             data_1960,
             size = 2) +
  geom_point(aes(Fertility, Life_exp, color = 'cyan'), alpha = 0.5,
             data_2015,
             size = 2) +
  ggtitle('Długość Życia a Płodność') +
  labs(y = 'Life Exp',
       color = NULL,
       shape = NULL) +
  scale_x_continuous(breaks = seq(0.5, 10, 0.5)) +
  scale_y_continuous(breaks = seq(20, 100, 5)) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
```

```
        legend.text = element_text(size = 15))
```



W obydwu przypadkach widzimy tu trend malejący, Jednak w krajach o wysokiej dzietności widzimy, że oczekiwana długość jest bardzo różna, jest to spowodowane głównie faktem, że są to kraje nisko-rozwinięte.

Podsumowanie

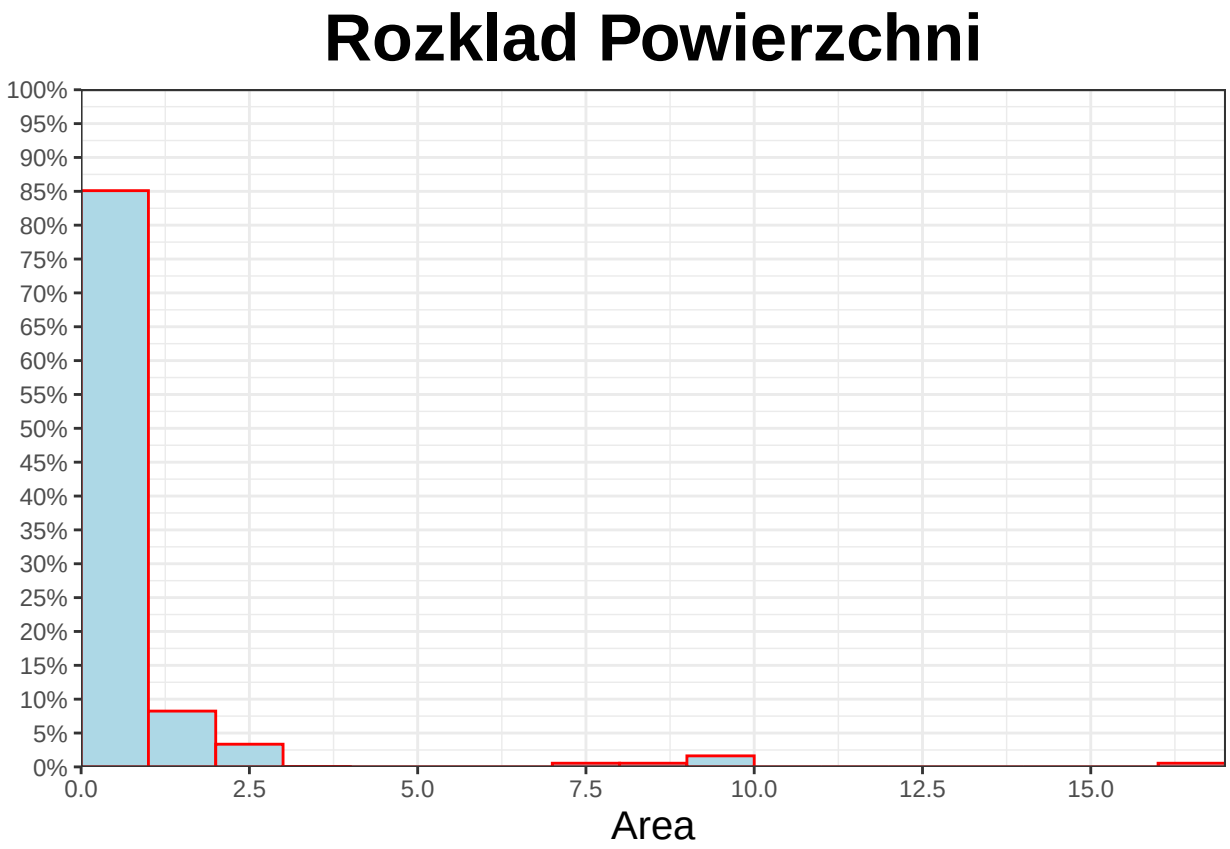
Widzimy więc, że mimo, że wartości obu zmiennych zmieniły się w czasie to zależność między nimi raczej pozostały podobne. Widzimy więc, że zależność wspomniane w hipotezie istnieje i raczej nie zniknie w miarę biegu czasu.

8. Gęstość Zaludnienia a Starzenie się Społeczeństwa

Na zakończenie naszego badania zajmijmy się zależnością między **powierzchnią**, a **procentem ludności powyżej 65 roku życia**. Powinniśmy się najpierw zająć rozkładem zmiennej **Area**.

Przystąpmy więc do zobrazowania dystrybucji zmiennej **Area** za pomocą histogramu.

```
ggplot(data) +
  geom_histogram(aes(Area, ..count../sum(..count..)),
                 boundary = 0,
                 binwidth = 1,
                 fill = 'light blue', color = 'red') +
  ggtitle('Rozkład Powierzchni') +
  expand_limits(x = 0, y = c(0, 1)) +
  scale_x_continuous(expand = c(0, NA), breaks = seq(0, 20, 2.5)) +
  scale_y_continuous(breaks = seq(0, 1, 0.05), expand = c(0, NA),
                     labels = scales::percent) +
  labs(y = NULL) +
  theme_bw() +
  theme(plot.title = element_text(size = 25, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15))
```



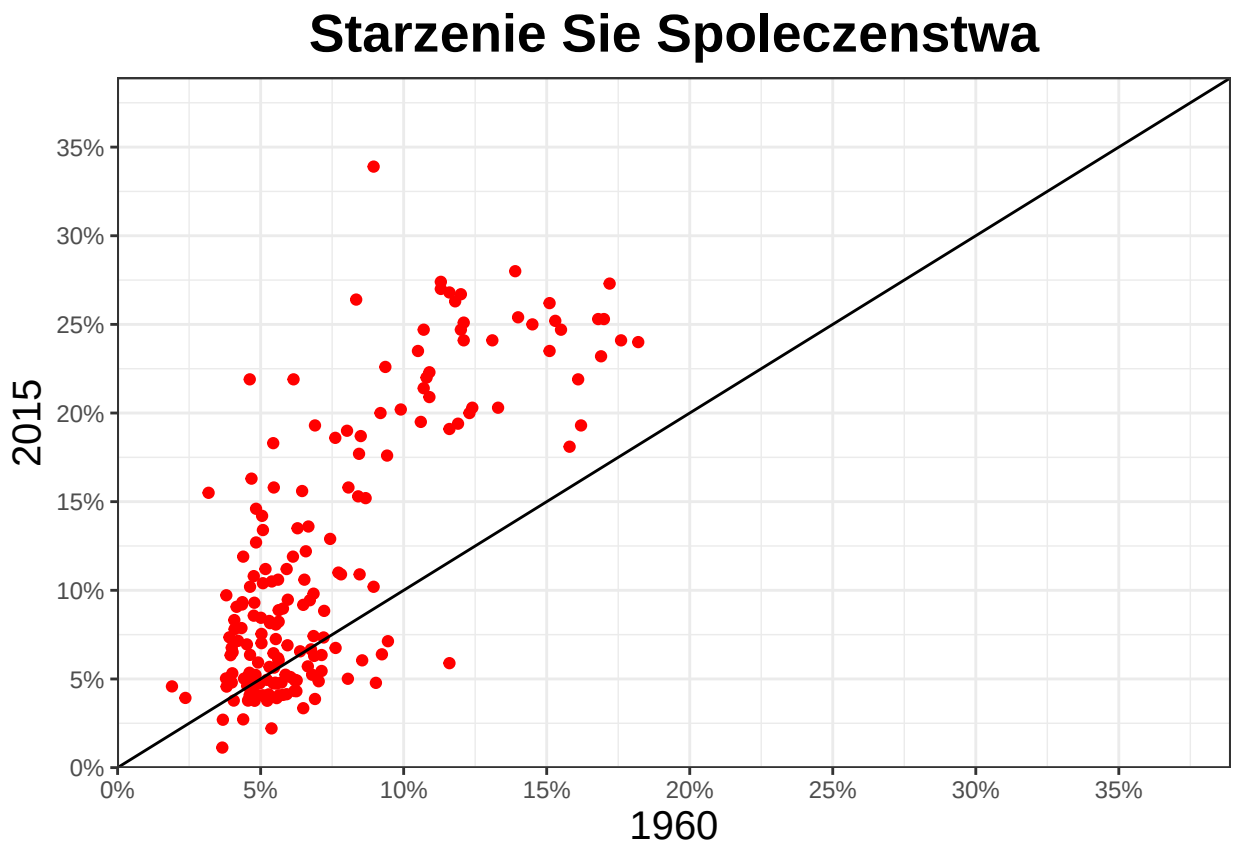
Możemy więc wnioskować, że rozsądne będzie zawężenie badanych wartości zmiennej **Area** do przedziału [0, 2], gdyż przedział ten będzie zawierać ponad 95% wartości zmiennych. Pozostałe 5% uznajmujemy za zbyt małą proporcją do wykonania satysfakcjonującej analizy.

Rozważmy także zmienność **procentowego udziału osób starych** na przestrzeni lat. Zobrazujmy tą zależność za pomocą lat skrajnych w naszym zestawie danych oraz wykresu punktowego.

```
data %>%
  select(Country, Year, Population_over_60) %>%
  mutate(Population_over_60 = Population_over_60 / 100) %>%
  spread(key = Year, value = Population_over_60) %>%
ggplot(aes_(as.name('1960'), as.name('2015'))) +
  geom_point(color = 'red') +
  geom_abline(slope = 1, intercept = 0) +
  ggtitle('Starzenie Się Społeczeństwa') +
  expand_limits(x = c(0, max(data$Population_over_60 / 100) + 0.05),
                y = c(0, max(data$Population_over_60 / 100) + 0.05)) +
  scale_x_continuous(breaks = seq(0, 1, 0.05), expand = c(0, NA),
                     labels = scales::percent) +
  scale_y_continuous(breaks = seq(0, 1, 0.05), expand = c(0, NA),
                     labels = scales::percent) +
  theme_bw() +
  theme(plot.title = element_text(size = 20, face = 'bold', hjust = 0.5),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15))
```

```
## Warning: `aes_()` was deprecated in ggplot2 3.0.0.
```

```
## i Please use tidy evaluation idioms with `aes()`
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

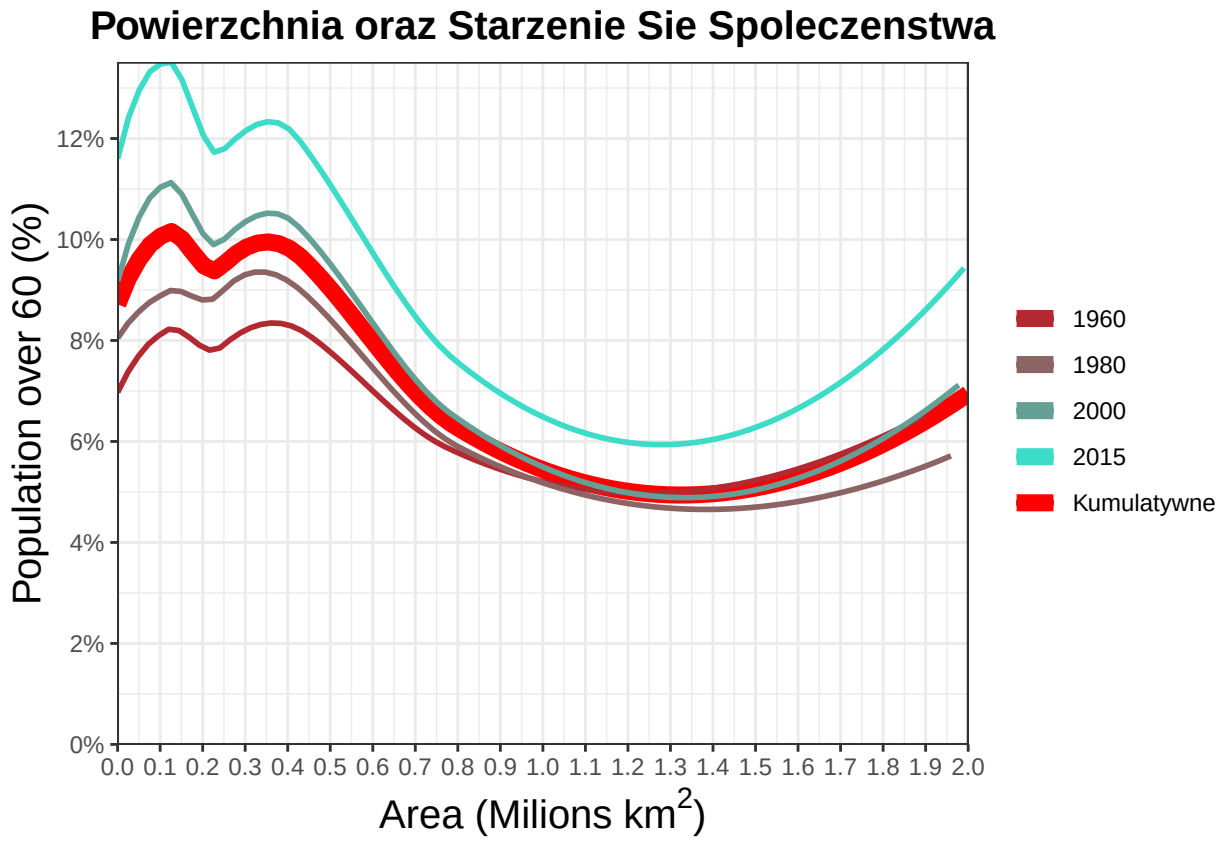


Widzimy tutaj, że społeczeństwo ewidentnie się starzeje, co jest związane ze zmiejszoną dzietnością oraz zwiększoną długością życia. Możemy się spodziewać, że tam gdzie odsetek osób starszych się zmniejszył to będę kraje słabo rozwinięte, szczególnie afrykańskie w których w tym czasie rozwinął się Aids.

Zbadajmy teraz zależność między tymi zmiennymi. Użyjemy metody powszechnie stosowanej już w tej pracy tzn. ekstrację trendu.

```
ggplot() +
  geom_smooth(aes(Area, Population_over_60 / 100,
                  color = 'Kumulatywne'),
              data %>%
                filter(Area <= 2), se = F,
              method = 'loess', formula = y ~ x, size = 3) +
  geom_smooth(aes(Area, Population_over_60 / 100,
                  color = '1960'), data_1960 %>%
                filter(Area <= 2), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Area, Population_over_60 / 100,
                  color = '1980'), data_1980 %>%
                filter(Area <= 2), se = F,
              method = 'loess', formula = y ~ x) +
  geom_smooth(aes(Area, Population_over_60 / 100,
```

```
                color = '2000'), data_2000 %>%
              filter(Area <= 2), se = F,
            method = 'loess', formula = y ~ x) +
geom_smooth(aes(Area, Population_over_60 / 100,
                color = '2015'), data_2015 %>%
              filter(Area <= 2), se = F,
            method = 'loess', formula = y ~ x) +
ggtitle('Powierzchnia oraz Starzenie Się Społeczeństwa') +
labs(x = TeX('Area (Milions $km^2$)'),
     y = 'Population over 60 (%)', parse = T,
     color = NULL) +
expand_limits(x = 0, y = 0) +
scale_x_continuous(expand = c(0, NA),breaks = seq(0, 2, 0.1)) +
scale_y_continuous(expand = c(0, NA), breaks = seq(0, 1, 0.02),
                   labels = scales::percent) +
theme_bw() +
theme(plot.title = element_text(size = 15, face = 'bold', hjust = 0.5),
      axis.title.x = element_text(size = 15),
      axis.title.y = element_text(size = 15)) +
scale_color_manual(values = colors)
```



Widzimy tutaj ciekawą zależność, która nie jest do końca jednoznaczna. Kraje o niskiej powierzchni są generalnie społeczeństwami starami, jednak trend nie jest jednoznacznie malejący, gdyż kraje o bardzo wysokiej powierzchni są generalnie
bardziej rozwinięte niż te o umiarkowanie dużej powierzchni.

Podsumowanie

Zależność jaką otrzymaliśmy jest praktycznie odwrotna do spodziewanej. Generalnie najbardziej "stare" kraje mają dosyć niską powierzchnią. Po głębszym przemyśleniu może to mieć sens, gdyż w tym przedziale znajduje się większość krajów europejskich. Analizę krajów o bardzo wysokiej powierzchnii pomijamy w tej pracy.

Zakończenie

Udało nam się zbadać wiele zależności między różnymi cechami społeczeństw. Nie wszystkie wyniki jakie otrzymaliśmy były takie jakich się spodziewaliśmy, jednak udało się nam potwierdzić dużą część naszych hipotez.

Mam nadzieję, że praca ta zaciekawiła oraz zinspirowała czytelnika do dalszego poszukiwania informacji o zależnościach pomiędzy różnymi cechami społeczeństw oraz krajów.