Uniwersytet Gdański

Wydział Matematyki, Fizyki i Informatyki

Instytut Informatyki

Grupowanie na podstawie bazy danych klientów

Maciej Witkowski

Nr albumu: 266722

Gdańsk, 19.05.2021

SPIS TREŚCI

WPROWADZENIE **1**

1. WSTĘPNA ANALIZA DANYCH **1**

1.1 Histogramy – wiek, roczny dochód, ocena wydawania 1

1.2 Wykres liczby kobiet i mężczyzn – porównanie 5

1.3 Wykres zależności pomiędzy kolumnami 7

1.4 Przygotowanie bazy danych do procesu klasteryzacji 14

1.5 Macierz kowariancji i korelacji 17

2. KLASTERYZACJA PODSTAWOWA **22**

2.1 Klasteryzacja z użyciem algorytmu k-średnich dla k=5 22

2.2 Wynik klasteryzacji na wykresie 2D 22

3. KLASTERYZACJA Z PCA **24**

3.1 Klasteryzacja z użyciem algorytmu k-średnich dla k=5 24

3.2 Wynik klasteryzacji na wykresie 2D 25

4. ANALIZA WYNIKÓW I WNIOSKI **27**

4.1 Porównanie klasteryzacji z użyciem PCA i bez 27

4.2 Opis i ocena potencjału każdej z grup 27

4.3 Kwestia dominacji płci w grupach 27

4.4 Znaczenie oraz wpływ zależności między kolumnami na wynik klasteryzacji 29

Gdańsk, 19.05.2021

WPROWADZENIE

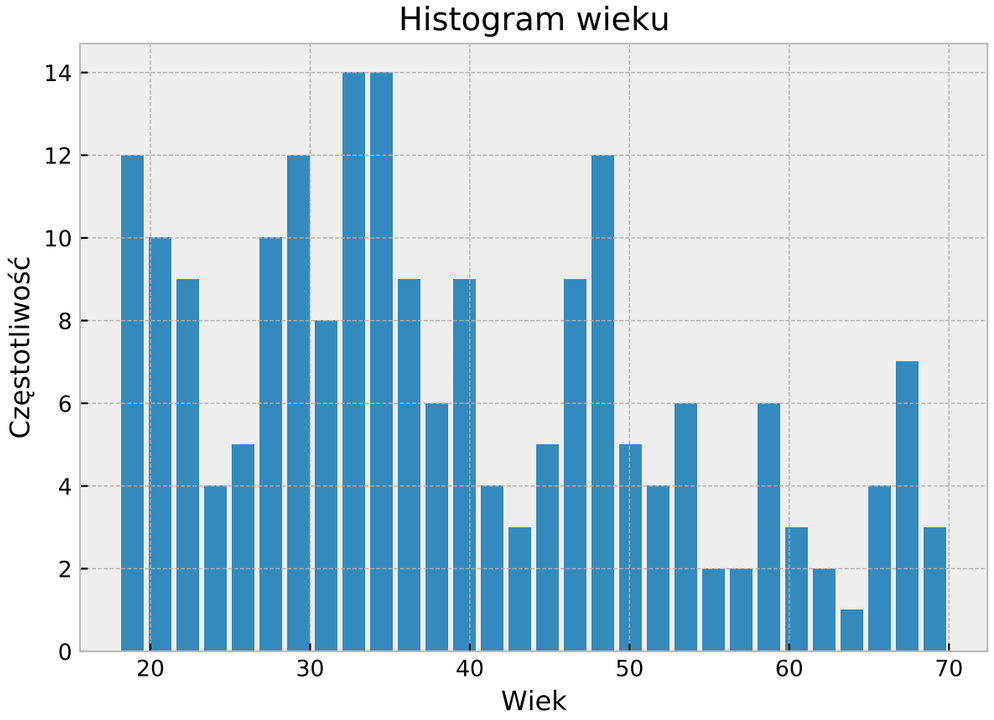
Celem projektu jest analiza klasteryzacji bazy danych Mall_Customers.csv, która zawiera dane 200 klientów centrum handlowego, za pomocą metody k-średnich przy użyciu algorytmu PCA i bez. Analizie zostanie poddana ocena potencjału każdej grupy pod względem przyszłych wyników sprzedaży, której wnioski pozwolą oszacować do kogo należy kierować najwięcej informacji marketingowych. Zostaną także zbadane takie aspekty jak wpływ i dominacja płci w każdej z poszczególnych grup oraz zależności między kolumnami w bazie danych, które mogą znacząco rzutować na wynik klasteryzacji.

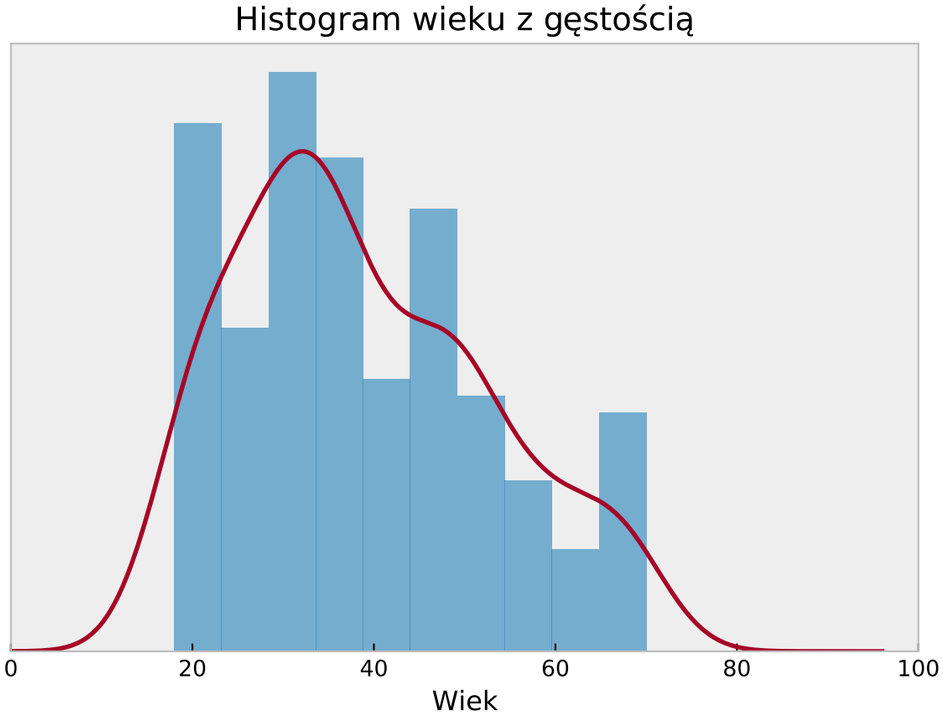
1. WSTĘPNA ANALIZA DANYCH

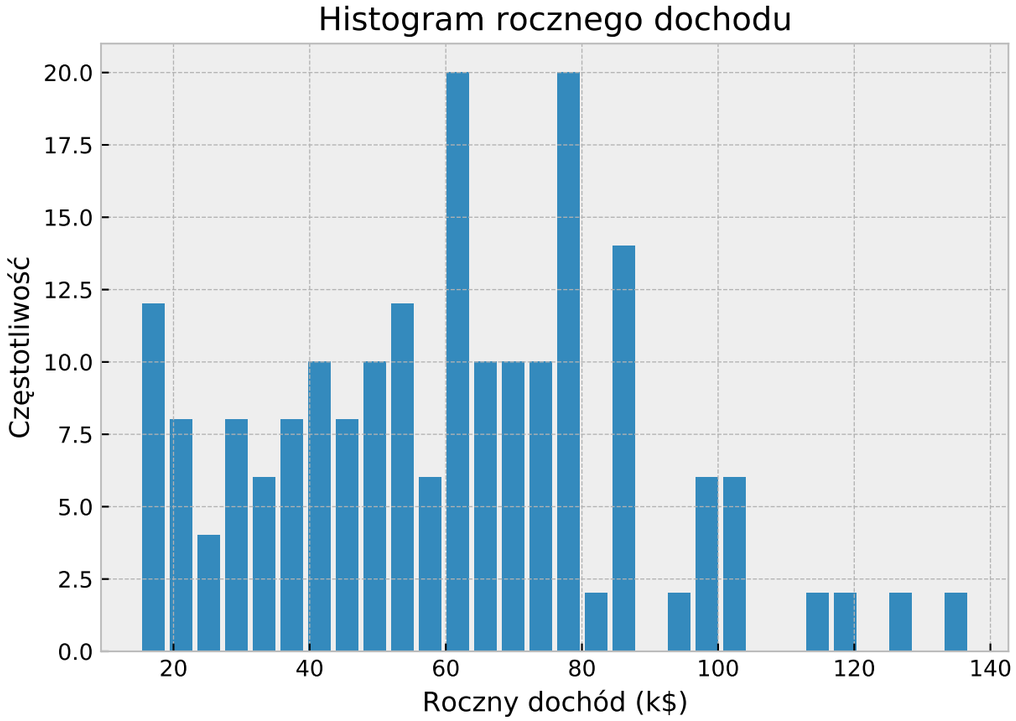
1.1. Histogramy – wiek, roczny dochód, ocena wydawania

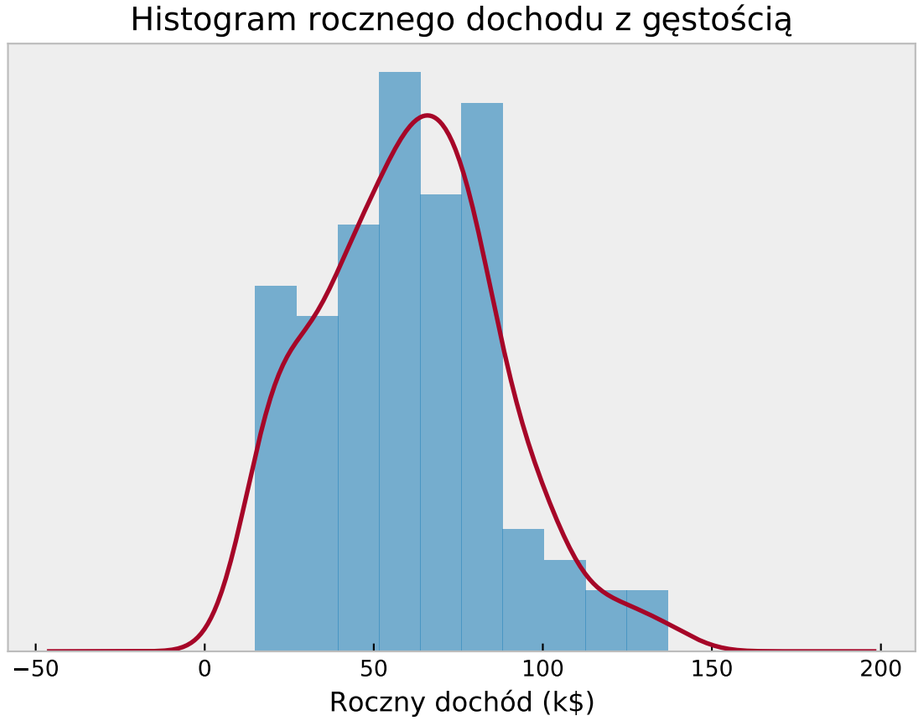
Na początku zostaną wygenerowane histogramy dla wieku, rocznego dochodu i oceny wydawania w celu wstępnego rozpoznania danych znajdujących się w bazie.

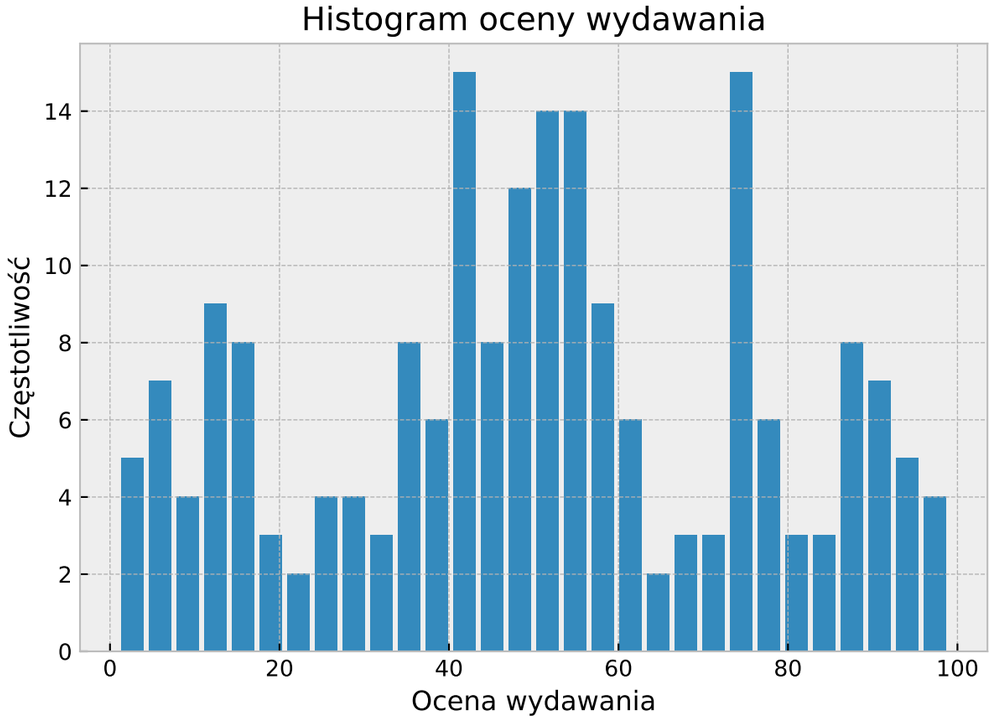
Gdańsk, 19.05.2021

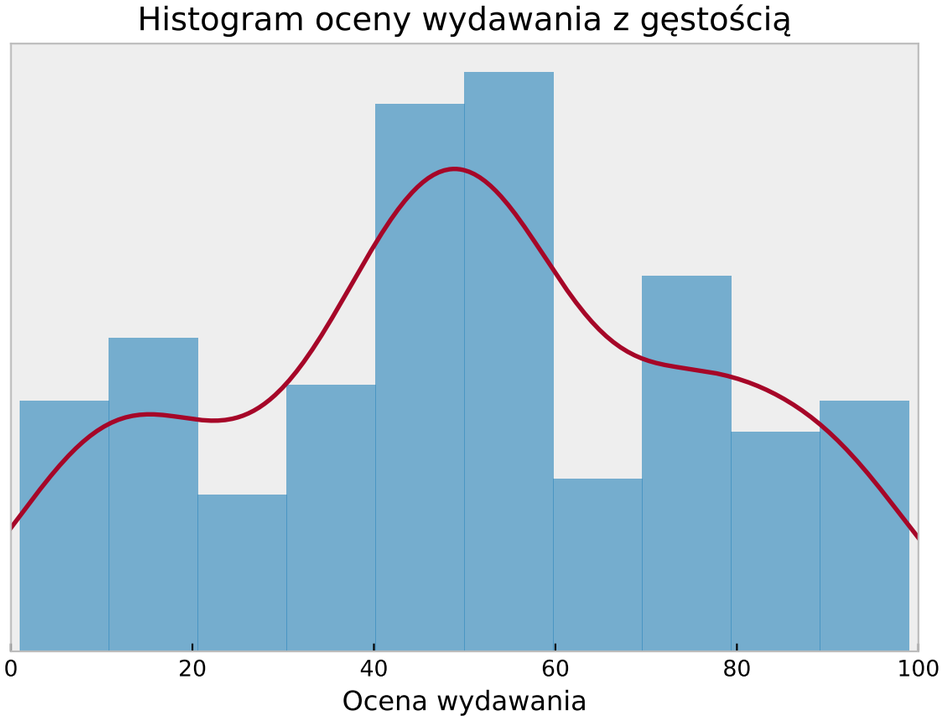

Histogram wieku
Częstotliwość
Wiek
0
2
4
6
8
10
12
14
20
30
40
50
60
70


Histogram wieku z gęstością
Wiek
0
20
40
60
80
100


Histogram rocznego dochodu
Częstotliwość
Roczny dochód (k$)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
20
40
60
80
100
120
140


Histogram rocznego dochodu z gęstością
Roczny dochód (k$)
−50
0
50
100
150
200

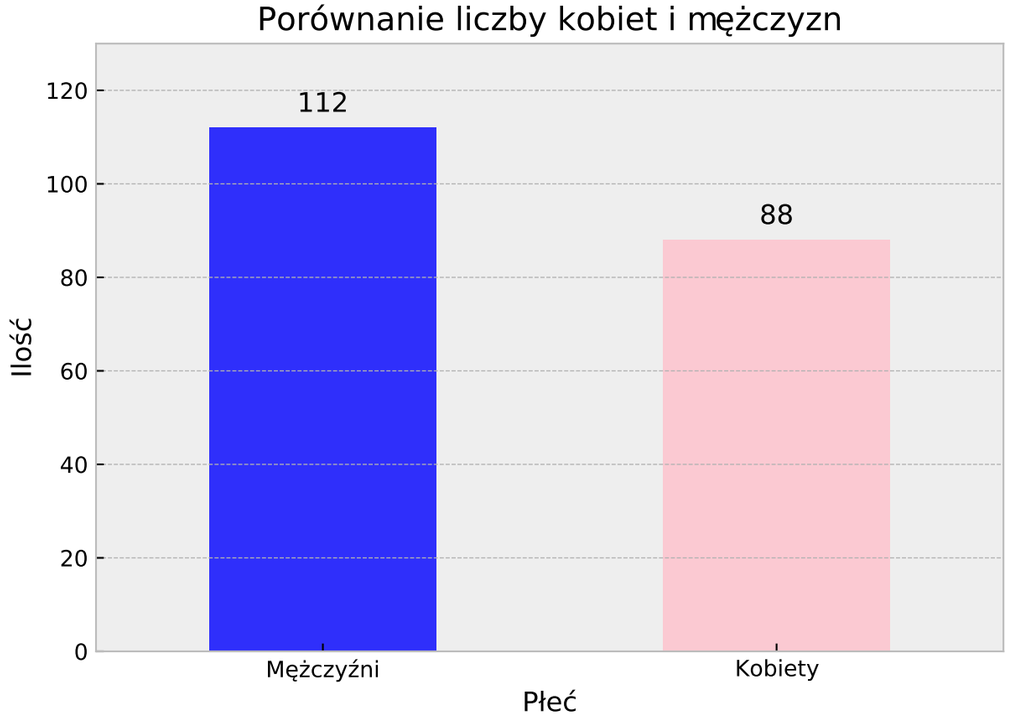

Histogram oceny wydawania
Częstotliwość
Ocena wydawania
0
2
4
6
8
10
12
14
0
20
40
60
80
100

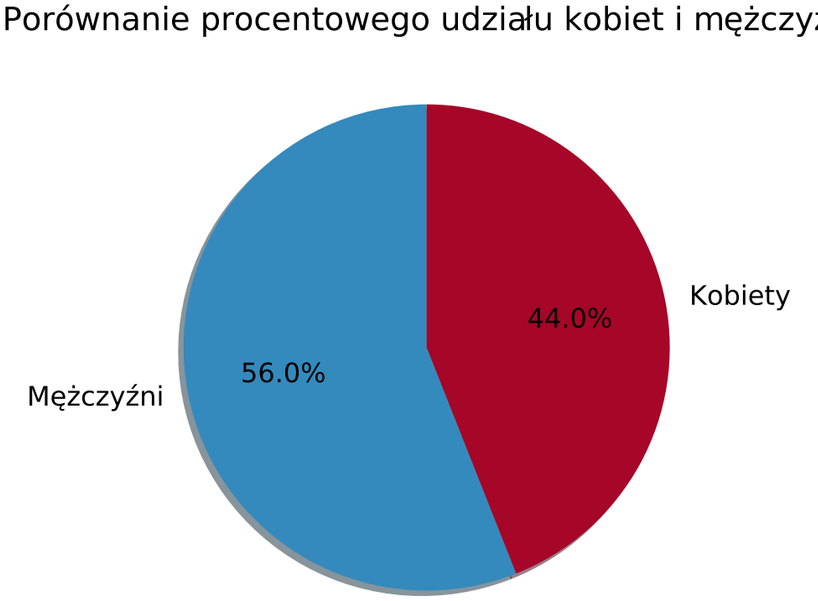

Histogram oceny wydawania z gęstością
Ocena wydawania
0
20
40
60
80
100

1.2 Wykres liczby kobiet i mężczyzn – porównanie

Następnie zbadany zostanie udział kobiet i mężczyzn, co pozwoli w przyszłości zweryfikować poprawność podziału na płcie w utworzonych w wyniku klasteryzacji grupach.

Gdańsk, 19.05.2021


Porównanie liczby kobiet i mężczyzn
120
100
80
60
40
20
0
Ilość
112
88
Mężczyźni
Kobiety
Płeć

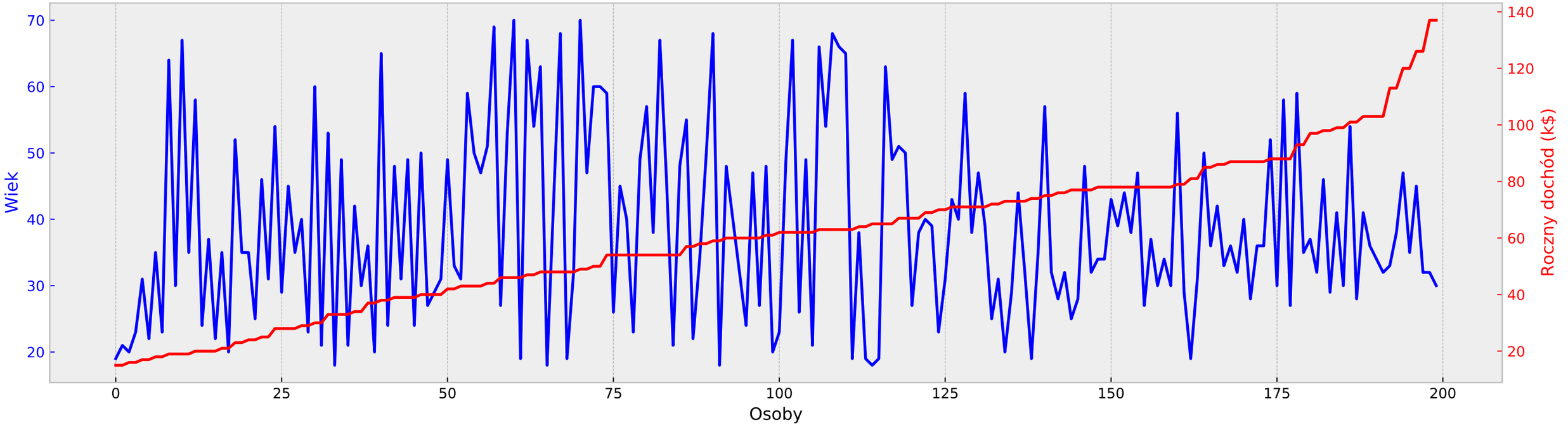

Porównanie procentowego udziału kobiet i mężczyzn
Kobiety
44.0%
56.0%
Mężczyźni

1.3 Wykresy zależności pomiędzy kolumnami

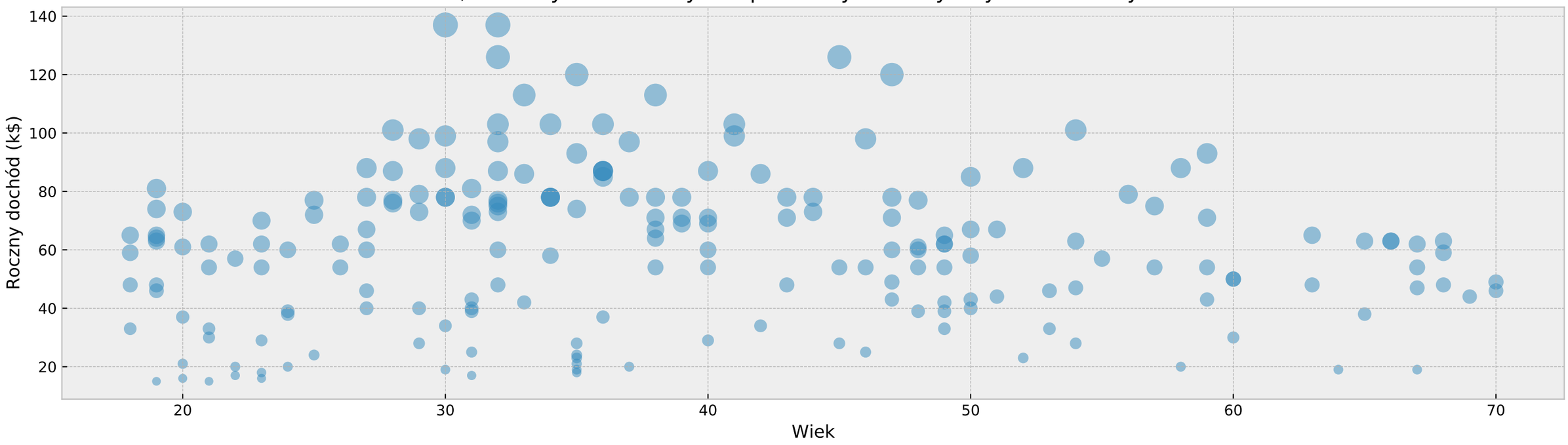
Na koniec zostaną wykazane zależności pomiędzy kolumnami, których wpływ zostanie oceniony w późniejszej analizie.

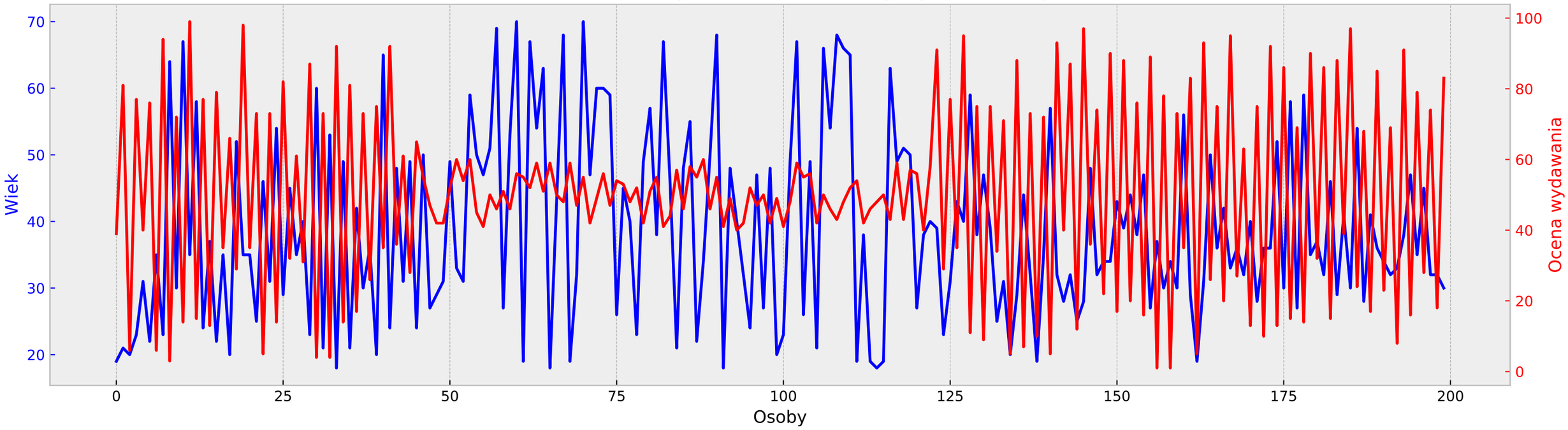
Gdańsk, 19.05.2021

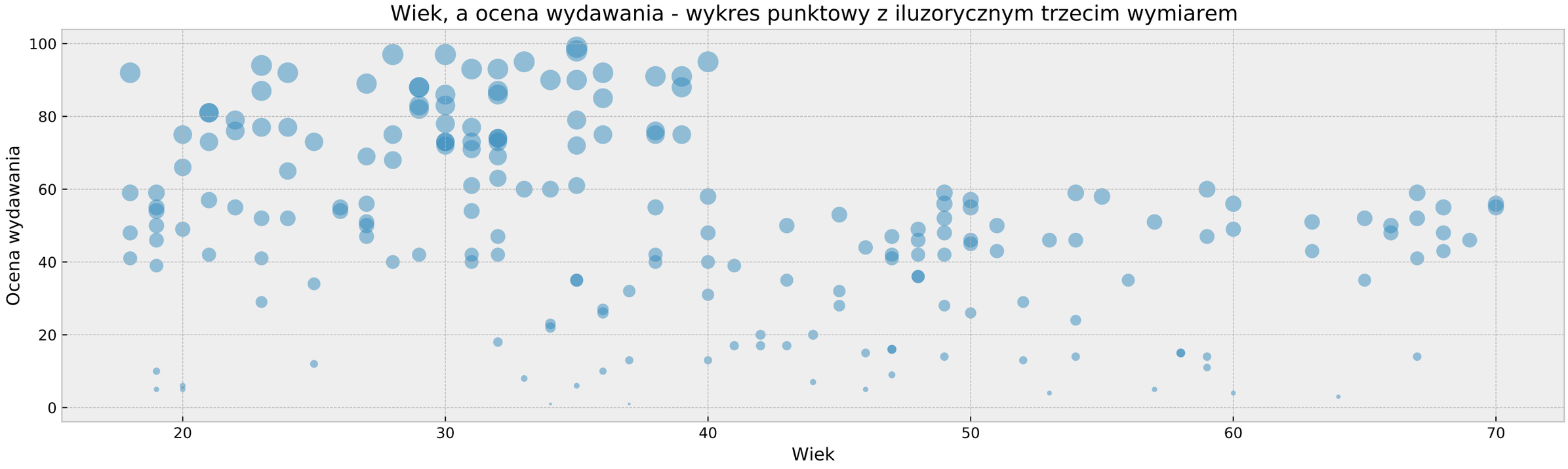
Wiek, a roczny dochód - bliźniaczy wykres liniowy



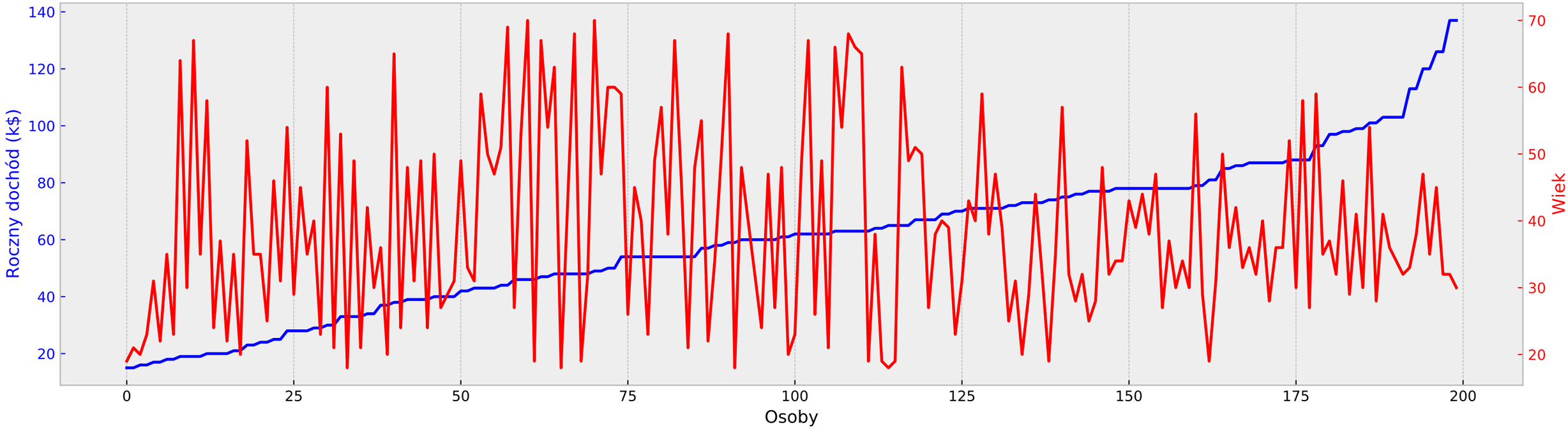
Wiek, a roczny dochód - wykres punktowy z iluzorycznym trzecim wymiarem



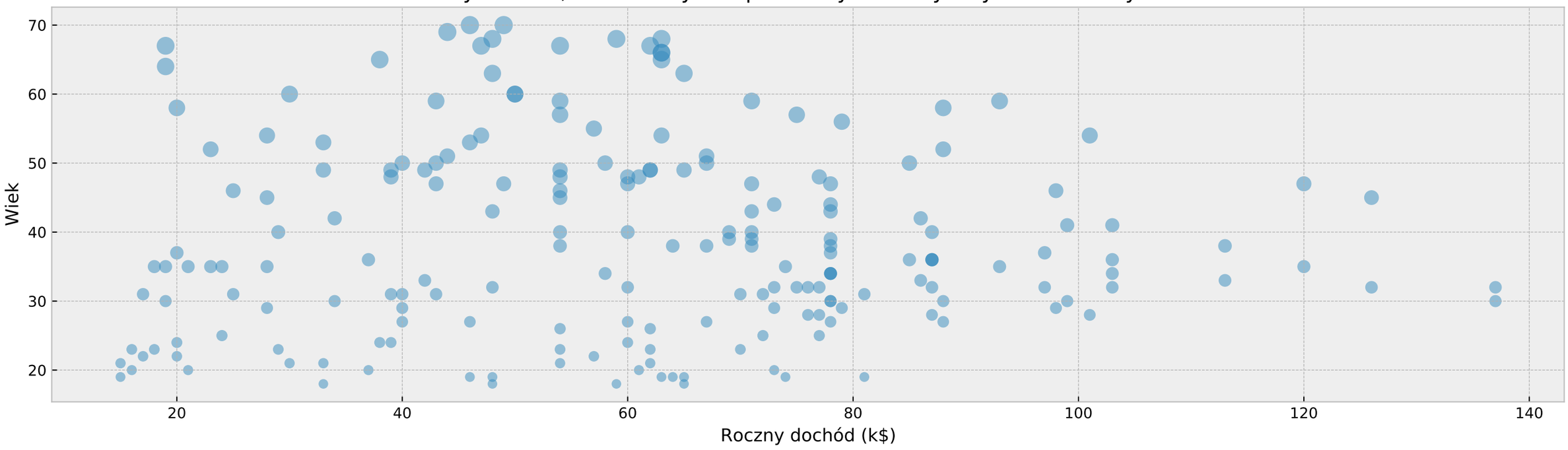

Wiek, a ocena wydawania - bliźniaczy wykres liniowy
Wiek
Ocena wydawania
Osoby
70
60
50
40
30
20
100
80
60
40
20
0
0
25
50
75
100
125
150
175
200

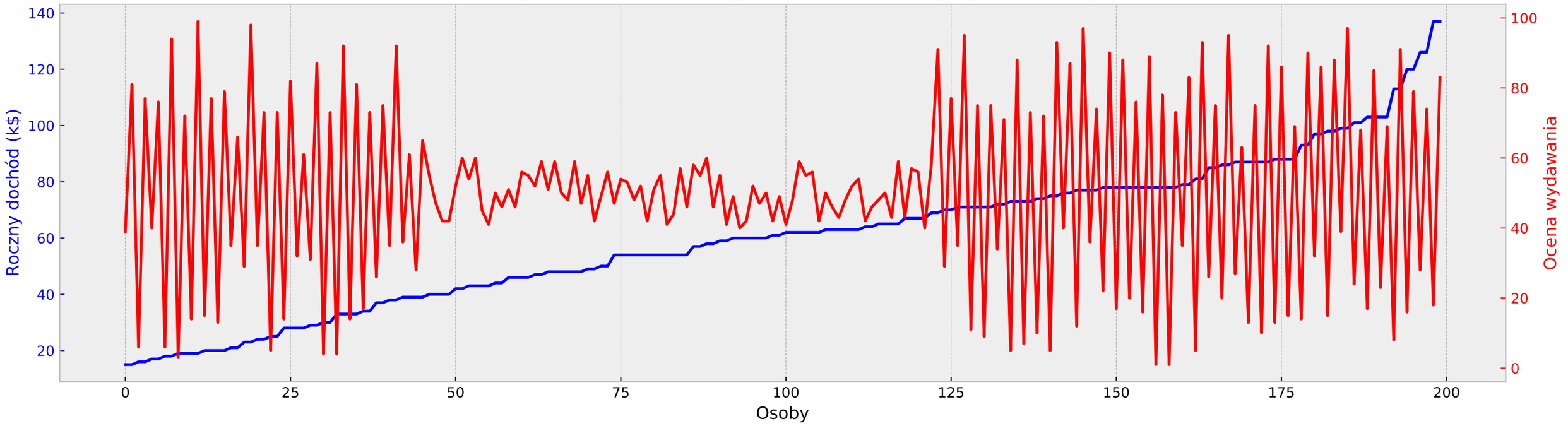

Wiek, a ocena wydawania - wykres punktowy z iluzorycznym trzecim wymiarem
Ocena wydawania
Wiek
100
80
60
40
20
0
20
30
40
50
60
70

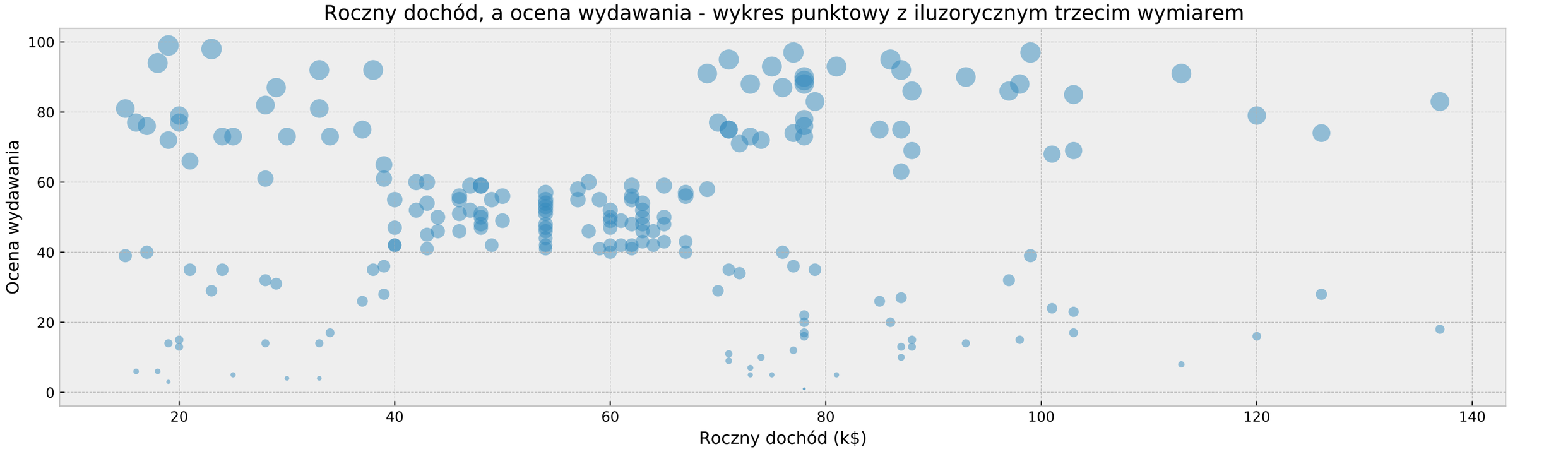
Roczny dochód, a wiek - bliźniaczy wykres liniowy

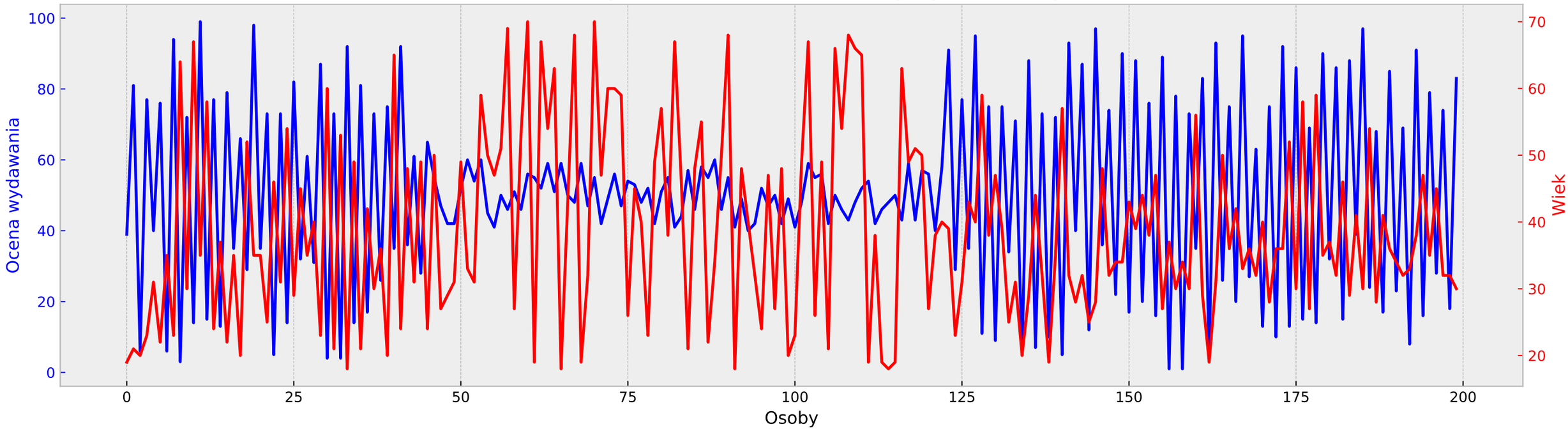


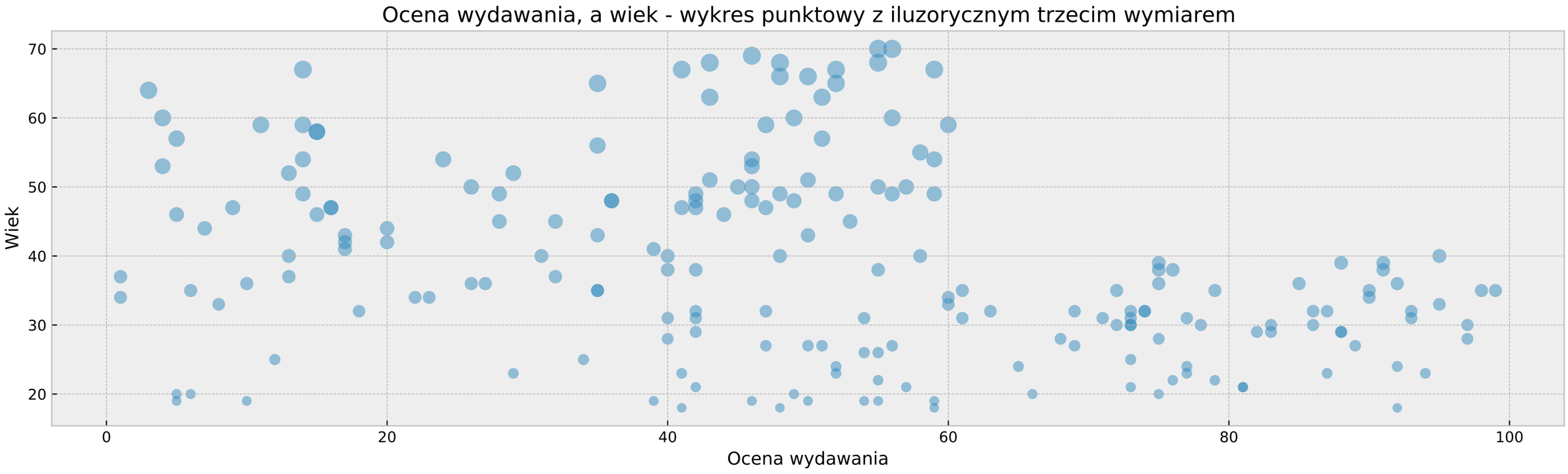
Roczny dochód, a wiek - wykres punktowy z iluzorycznym trzecim wymiarem

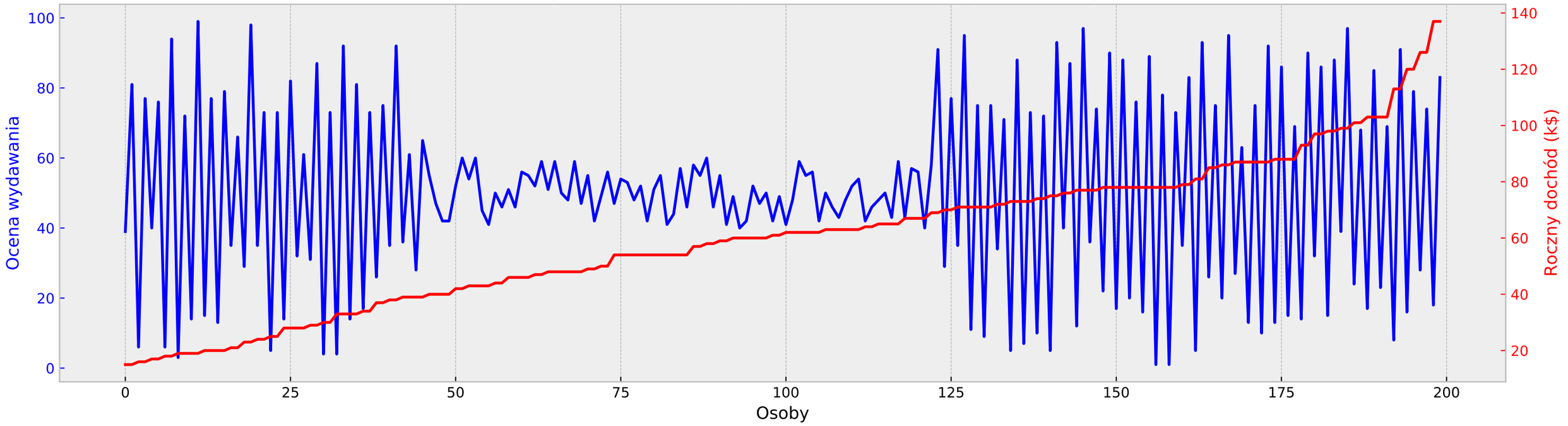


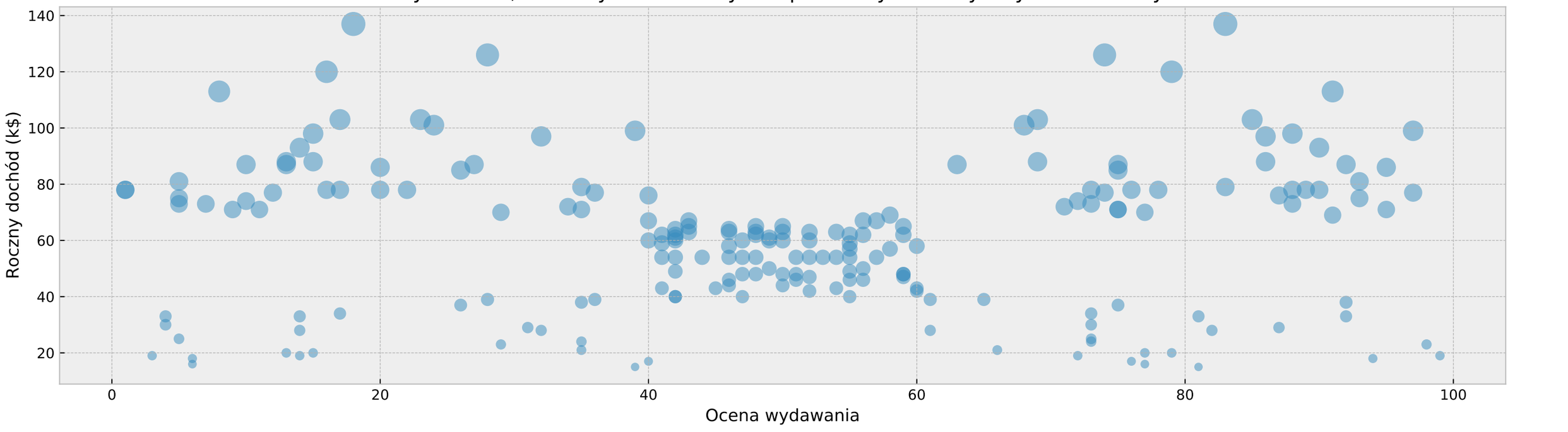

Roczny dochód, a ocena wydawania - bliźniaczy wykres liniowy
Roczny dochód (k$)
Ocena wydawania
Osoby
140
120
100
80
60
40
20
0
25
50
75
100
125
150
175
200


Roczny dochód, a ocena wydawania - wykres punktowy z iluzorycznym trzecim wymiarem
Ocena wydawania
Roczny dochód (k$)
100
80
60
40
20
0
20
40
60
80
100
120
140


Ocena wydawania, a wiek - bliźniaczy wykres liniowy
Ocena wydawania
Wiek
Osoby
0
20
40
60
80
100
0
25
50
75
100
125
150
175
200
20
30
40
50
60
70


Ocena wydawania, a wiek - wykres punktowy z iluzorycznym trzecim wymiarem
Wiek
Ocena wydawania
20
30
40
50
60
70
0
20
40
60
80
100


Ocena wydawania, a roczny dochód - bliźniaczy wykres liniowy
Ocena wydawania
Roczny dochód (k$)
Osoby
0
20
40
60
80
100
120
140
0
25
50
75
100
125
150
175
200


Ocena wydawania, a roczny dochód - wykres punktowy z iluzorycznym trzecim wymiarem
Roczny dochód (k$)
Ocena wydawania
20
40
60
80
100
120
140
0
20
40
60
80
100

1.4 Przygotowanie bazy danych do procesu klasteryzacji

Analizowana baza danych potrzebuje niezbędnych zmian, zanim zostanie poddana procesowi klasteryzacji. Na początku została wyświetlona struktura danych, która pomoże w ocenie, czy dana kolumna potrzebuje modyfikacji, bądź normalizacji.

```
<-------------------------------------------------------->
200    0.5
63     0.5
      ...
1      0.5
Name: CustomerID, Length: 200, dtype: float64
<-------------------------------------------------------->

<-------------------------------------------------------->
Female    56.0
Male      44.0
Name: Gender, dtype: float64
<-------------------------------------------------------->

<-------------------------------------------------------->
32    5.5
35    4.5
19    4.0
30    3.5
27    3.0
34    2.5
      ...
54    2.0
60    1.5
41    1.0
44    1.0
56    0.5
Name: Age, dtype: float64
<-------------------------------------------------------->

<-------------------------------------------------------->
54     6.0
60     3.0
      ...
15     1.0
Name: Annual Income (k$), Length: 64, dtype: float64
<-------------------------------------------------------->

<-------------------------------------------------------->
42    4.0
55    3.5
46    3.0
     ...
99    0.5
Name: Spending Score (1-100), Length: 84, dtype: float64
<-------------------------------------------------------->
```

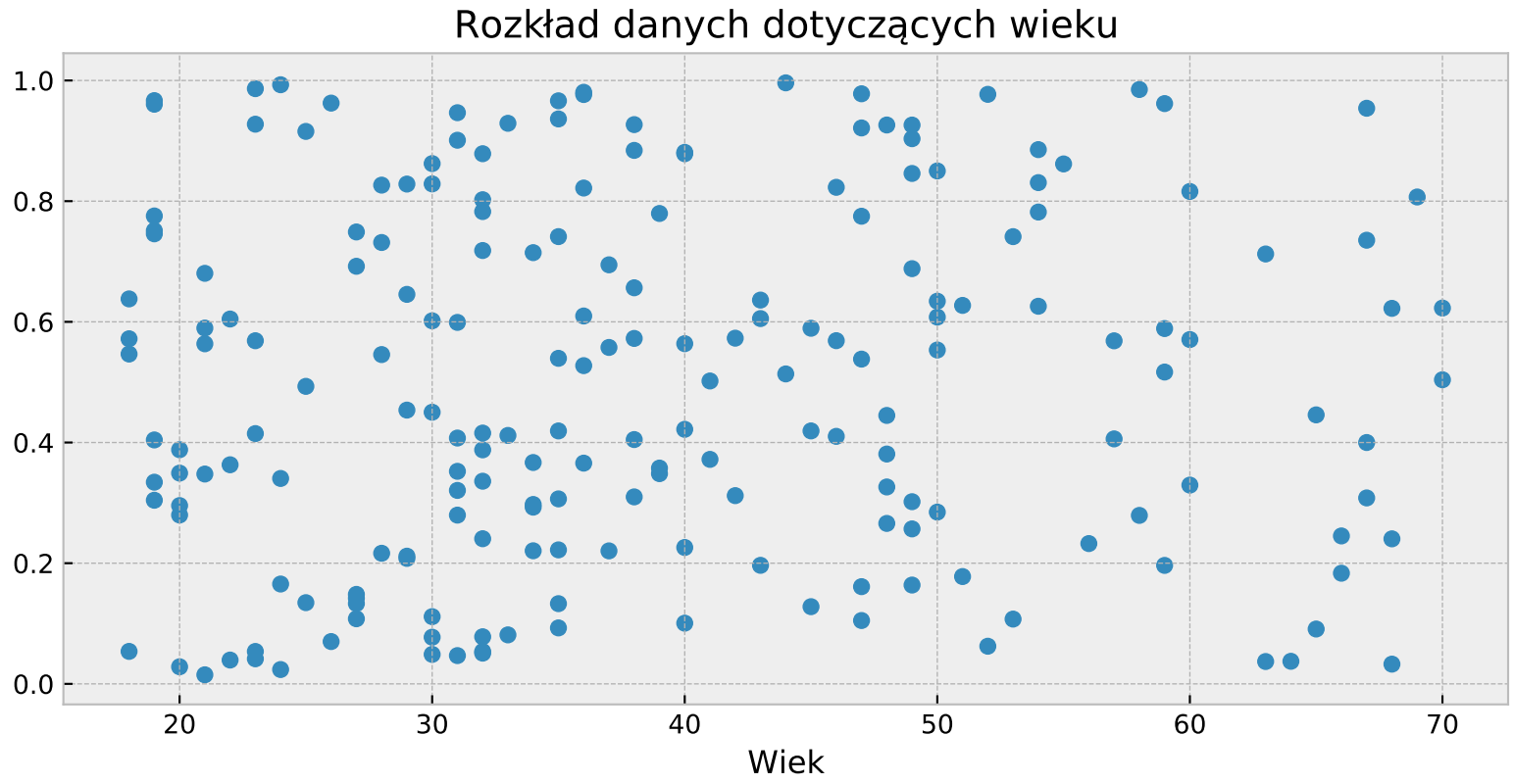
Gdańsk, 19.05.2021

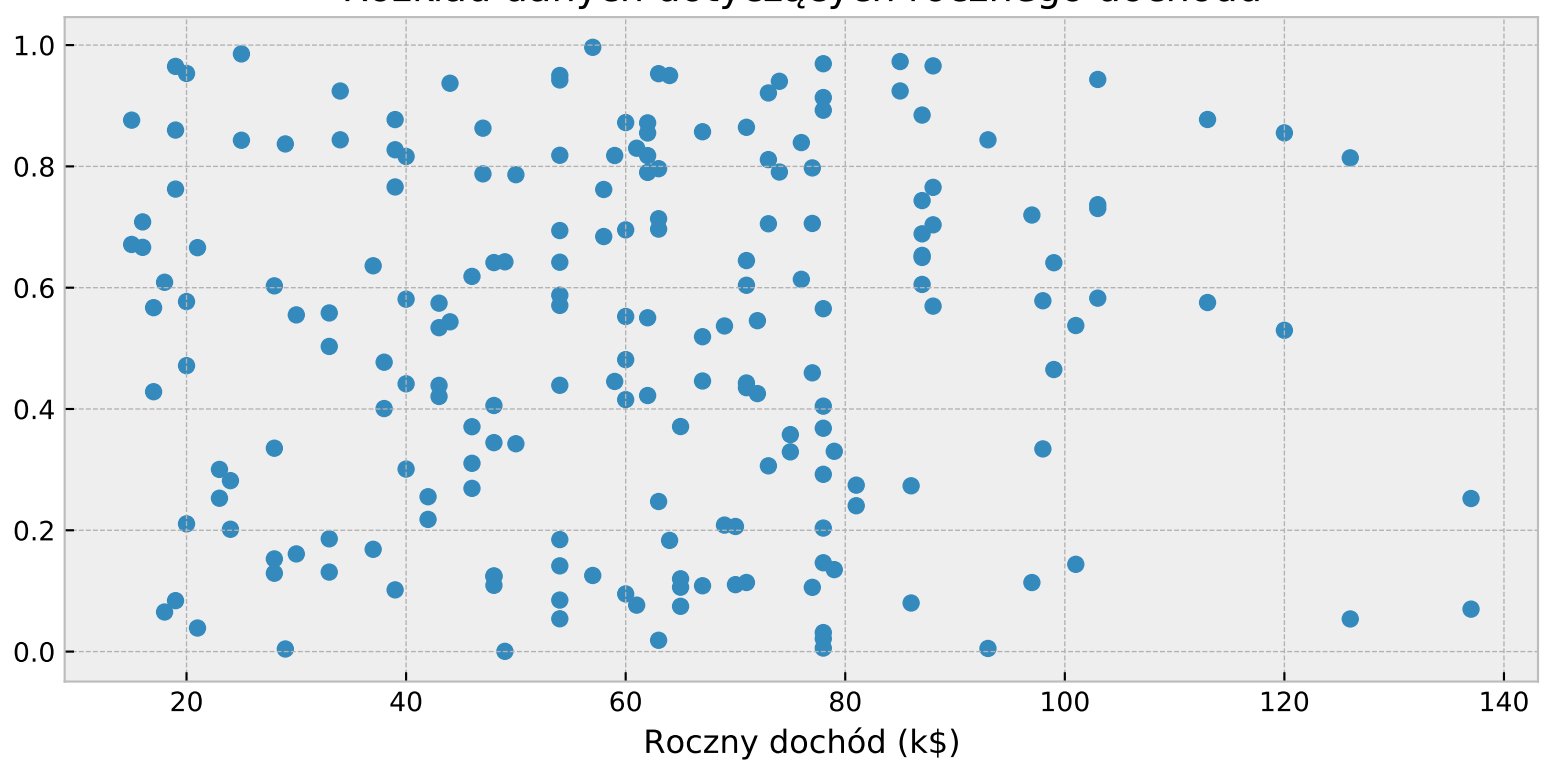
Na podstawie wyświetlonych informacji można od razu stwierdzić, że kolumna o nazwie CustomerID, która zawiera indeksy każdego rekordu, jest zbędna przy procesie klasteryzacji, a nawet może zaburzyć jej ocenę, dlatego została usunięta. Następnie została sprawdzona poprawność danych znajdujących się w bazie pod kątem błędnych wartości oraz pustych wierszy.

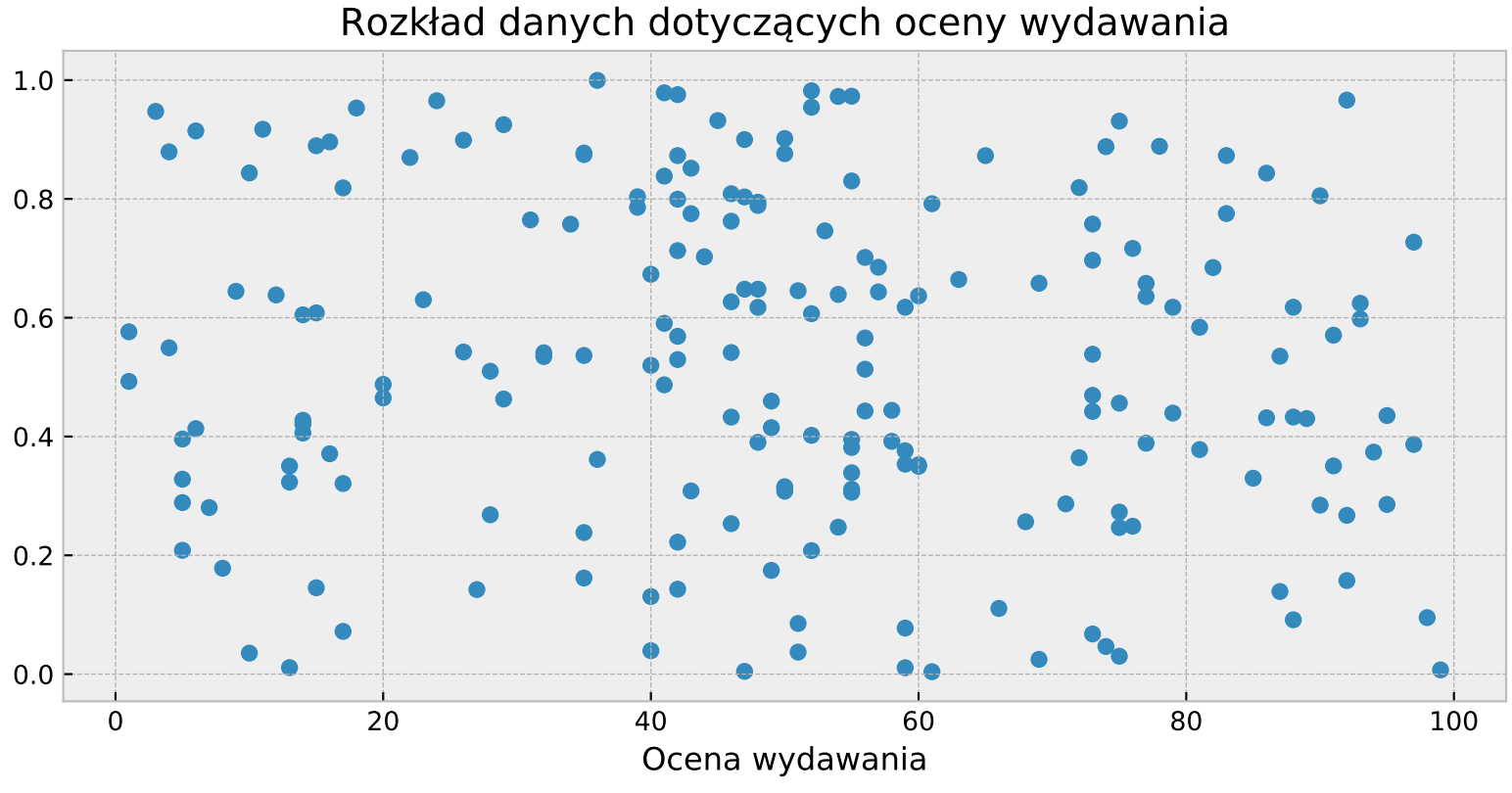
```
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Gender                  200 non-null    object
 1   Age                     200 non-null    int64
 2   Annual Income (k$)      200 non-null    int64
 3   Spending Score (1-100)  200 non-null    int64
dtypes: int64(3), object(1)
```

Dzięki powyższym informacjom, zauważono, że kolumna Gender, zawiera wartości nie liczbowe, dlatego została poddana normalizacji na wartości 0 lub 1. Na koniec sprawdzono rozkład danych w pozostałych kolumnach, w celu sprawdzenia, czy nie występują przypadki ekstremalne, które mogą przeszkodzić w poprawnej analizie grup po klasteryzacji.

Gdańsk, 19.05.2021


Rozkład danych dotyczących wieku
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
Wiek


Rozkład danych dotyczących rocznego dochodu
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
120
140
Roczny dochód (k$)


Rozkład danych dotyczących oceny wydawania
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Ocena wydawania

Na podstawie wykresów, można stwierdzić, że jedyną kolumną, który mogłaby by być poddana modyfikacji jest kolumna zawierająca roczny dochód, ze względu na niewielką ilość przypadków wykraczających poza przedział od 0 do 100. Niestety ze względu na wymagania projektu, nie zdecydowałem się na żadną modyfikację.

1.5 Macierz kowariancji i korelacji

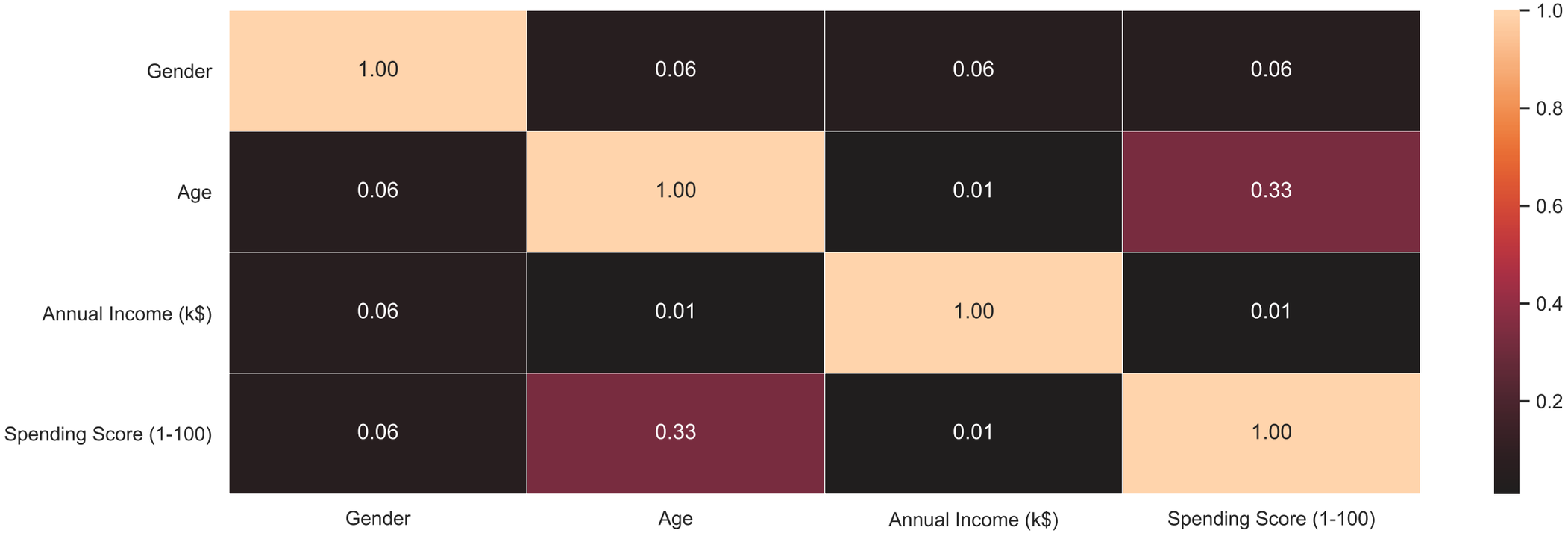
Wygenerowane zostały wykresy obrazujące macierz kowariancji i korelacji wraz z tabelami, które pomogą nam oszacować przydatność konkretnych kolumn w procesie klasteryzacji.

Gdańsk, 19.05.2021

	Gender	Age	Annual Income (k$)	Spending Score (1-100)
Gender	0.247638	0.423116	0.737286	-0.746734
Age	0.423116	195.133166	-4.548744	-118.040201
Annual Income (k$)	0.737286	-4.548744	689.835578	6.716583
Spending Score (1-100)	-0.746734	-118.040201	6.716583	666.854271


Gender
Age
Annual Income (k$)
Spending Score (1-100)
0.25
0.42
0.74
-0.75
0.42
195.13
-4.55
-118.04
0.74
-4.55
689.84
6.72
-0.75
-118.04
6.72
666.85
Gender
Age
Annual Income (k$)
Spending Score (1-100)
600
500
400
300
200
100
0
−100

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1.000000	0.026763	0.977548	0.013835
Age	0.026763	1.000000	0.012398	0.327227
Annual Income (k$)	0.977548	0.012398	1.000000	0.009903
Spending Score (1-100)	0.013835	0.327227	0.009903	1.000000


Gender
Age
Annual Income (k$)
Spending Score (1-100)
1.00
0.06
0.06
0.06
0.06
1.00
0.01
0.33
0.06
0.01
1.00
0.01
0.06
0.33
0.01
1.00
Gender
Age
Annual Income (k$)
Spending Score (1-100)
1.0
0.8
0.6
0.4
0.2

2. KLASTERYZACJA PODSTAWOWA

2.1 Klasteryzacja z użyciem algorytmu k-średnich dla k=5

Zgodnie z wymaganiami projektu, klasteryzacja ma zostać przeprowadzona na podstawie dwóch kolumn w bazie. Wyboru dokonałem na podstawie macierzy kowariancji i korelacji oraz wariancji, wartości własnej i wektora każdej z kolumn. Dzięki wysokim wartościom w macierzy kowariancji i korelacji, możemy zauważyć, że trzy kolumny zawierają w sobie dużo przydatnych do procesu klasteryzacji informacji, z czego dwie z tych trzech (roczny dochód i ocena wydawania) mają tych informacji najwięcej. Świadczy o tym wysoka wartość wariancji dla tychże kolumn.

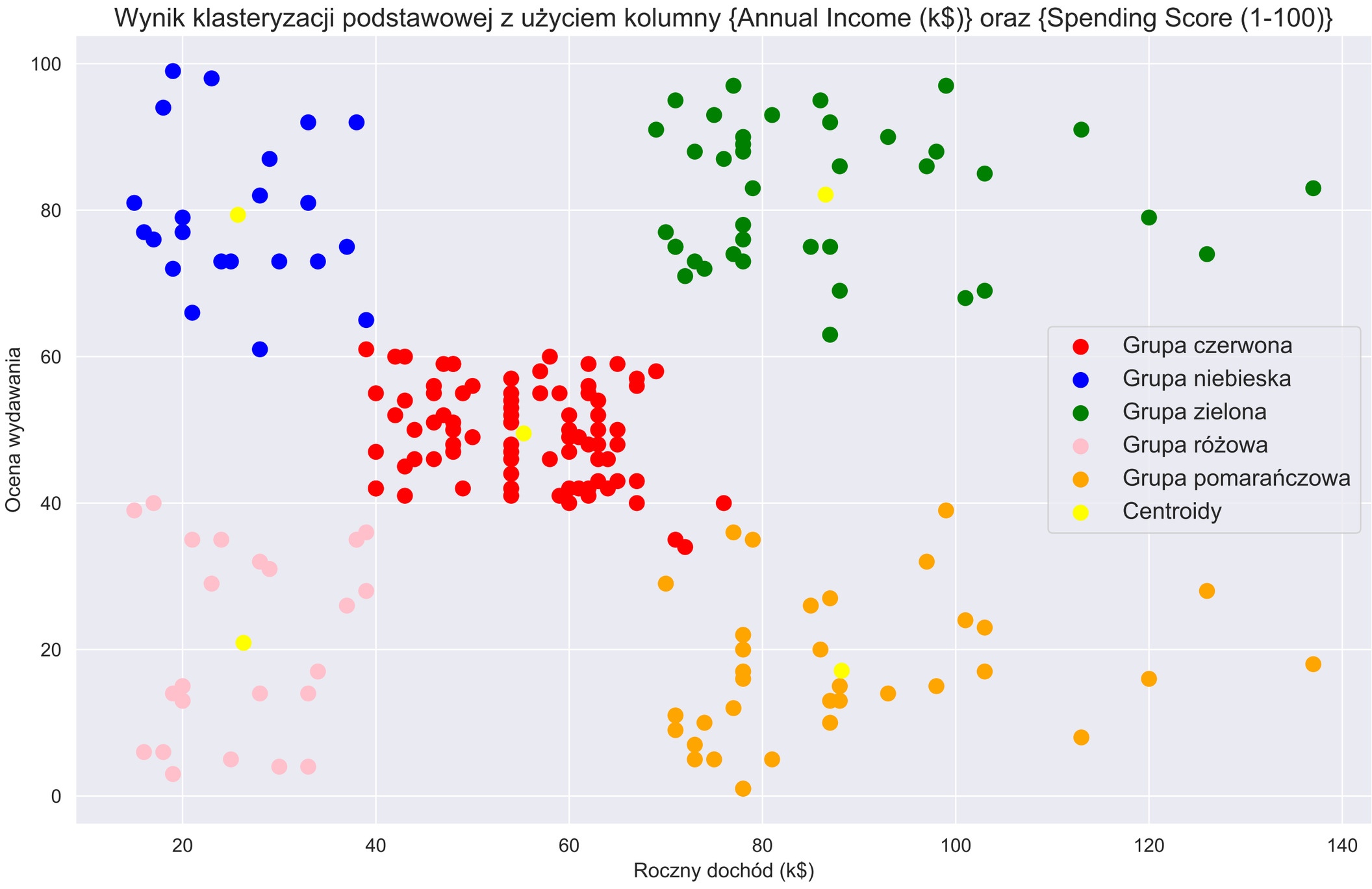
```
Kolumna: Gender
  Wariancja: 0.24763819095477396
  Wartość własna: 0.24549857825121535
  Wektor: [1.00 -0.00 -0.00 -0.00]

Kolumna: Age
  Wariancja: 195.1331658291457
  Wartość własna: 167.22888103525213
  Wektor: [-0.00 -0.97 -0.13 -0.19]

Kolumna: Annual Income (k$)
  Wariancja: 689.8355778894472
  Wartość własna: 684.3318413395426
  Wektor: [-0.00 -0.01 -0.81 0.59]

Kolumna: Spending Score (1-100)
  Wariancja: 666.8542713567839
  Wartość własna: 700.2644323132857
  Wektor: [0.00 -0.23 0.57 0.79]
```

Na podstawie wysokich wartości własnych (684.33 i 700.26), możemy z pewnością stwierdzić, że kolumny zawierające roczny dochód oraz ocenę wydawania, zawierają najwięcej informacji i to one będą najbardziej istotne dla procesu klasteryzacji.

2.2 Wynik klasteryzacji na wykresie 2D

Do przedstawienia wyników klasteryzacji z użyciem algorytmu k-średnich dla k=5 na podstawie kolumn Annual Income (k$) oraz Spending Score (1-100) posłużyłem się wykresem 2D. Wszystkie powstałe pięć grup zostały oznaczone na wykresie innym kolorem wraz z centroidami.

Gdańsk, 19.05.2021


Wynik klasteryzacji podstawowej z użyciem kolumny {Annual Income (k$)} oraz {Spending Score (1-100)}
Ocena wydawania
Roczny dochód (k$)
0
20
40
60
80
100
20
40
60
80
100
120
140
Grupa czerwona
Grupa niebieska
Grupa zielona
Grupa różowa
Grupa pomarańczowa
Centroidy

3. KLASTERYZACJA Z PCA

3.1 Klasteryzacja z użyciem algorytmu k-średnich dla k=5

Zgodnie z wymaganiami projektu, na początek ustaliłem liczbę wymiarów, do której ma zostać zredukowana baza danych. Do ustalenia tej wartości posłużyłem się wskaźnikiem poziomu wariancji (poziom starty informacji), który zgodnie z definicją nie powinien, być mniejszy niż 80%. Poniżej prezentuje wyniki poziomu wariancji dla każdej możliwej ilości wymiarów.

```
Liczba wymiarów: 1
  Wariacja każdego z wymiarów: [
         0.45118077
  ]
  Poziom wariancji (strata informacji): 45.12 %

Liczba wymiarów: 2
  Wariacja każdego z wymiarów: [
      0.45118077
      0.44091539
  ]
  Poziom wariancji (strata informacji): 89.21 %

Liczba wymiarów: 3
  Wariacja każdego z wymiarów: [
      0.45118077
      0.44091539
      0.10774566
  ]
  Poziom wariancji (strata informacji): 99.98 %

Liczba wymiarów: 4
  Wariacja każdego z wymiarów: [
      4.51180770e-01
      4.40915393e-01
      1.07745663e-01
      1.58174873e-04
  ]
  Poziom wariancji (strata informacji): 100.0 %
```

Na podstawie powyższych wyników, można przystąpić do redukcji bazy danych do dwóch wymiarów przy poziomie wariancji wynoszącym 89.21%. Z poprzednich obserwacji można wywnioskować, że tymi dwoma wybranymi kolumnami, będzie kolumna Annual Income (k$) oraz Spending Score (1-100), jednakże można tę informację zweryfikować na podstawie wektorów tychże kolumn. Poniższa tabela zawiera komponenty wybrane przez algorytm PCA, jak widać wektory w tej tabeli są zgodne z wektorami kolumny Annual Income (k$) oraz Spending Score (1-100).

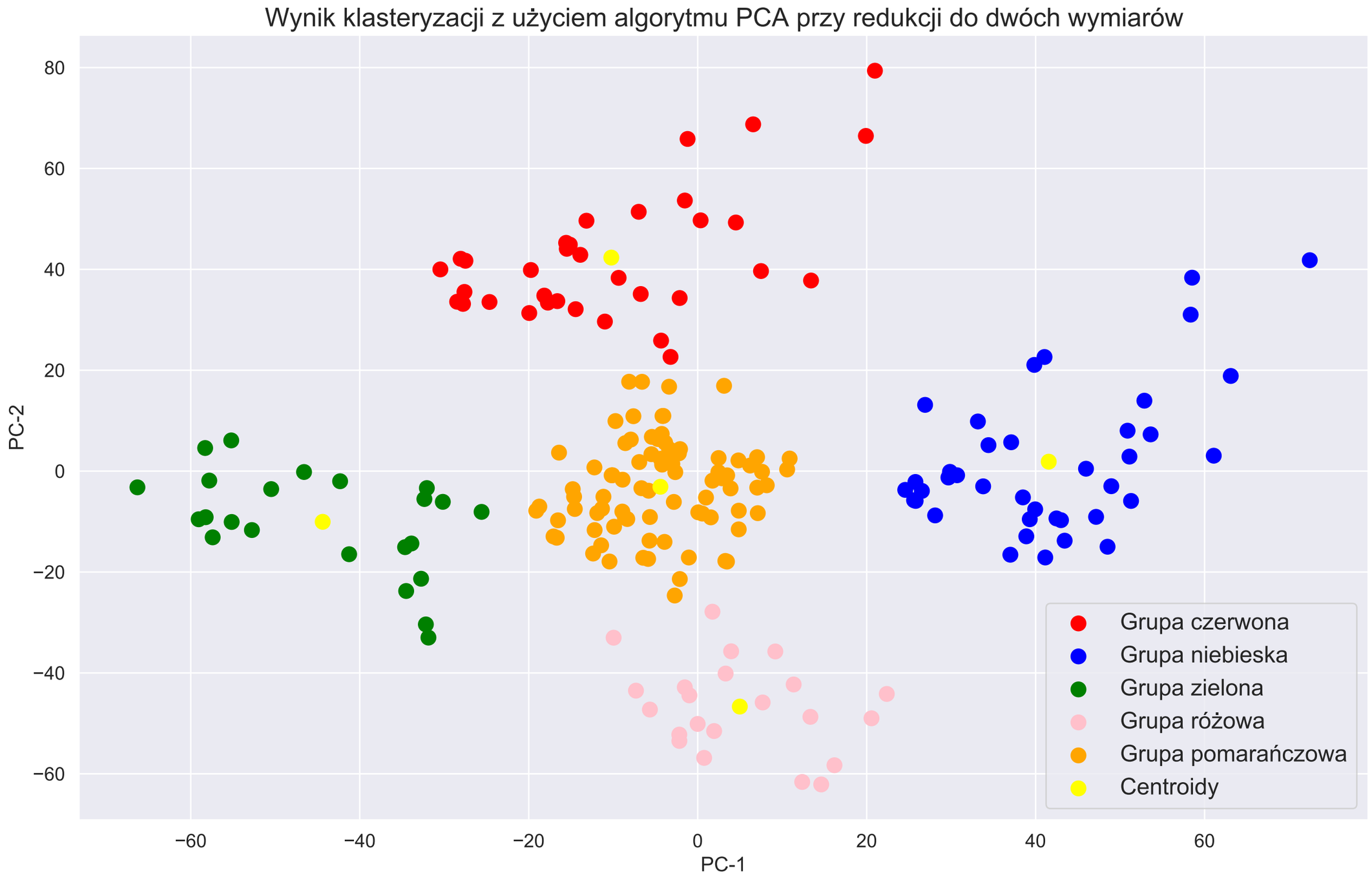
	Gender	Age	Annual Income (k$)	Spending Score (1-100)
PC-1	-0.000333	-0.188977	0.588623	0.786009
PC-2	0.001579	0.130961	0.808388	-0.573895

Gdańsk, 19.05.2021

3.1 Wynik klasteryzacji na wykresie 2D

Podobnie jak poprzednio do przedstawienia wyników klasteryzacji z użyciem algorytmu k-średnich dla k=5 przy ograniczeniu wielowymiarowości do 2 za pomocą algorytmu PCA posłużyłem się wykresem 2D. Wszystkie powstałe pięć grup zostały oznaczone na wykresie innym kolorem wraz z centroidami.

Gdańsk, 19.05.2021


Wynik klasteryzacji z użyciem algorytmu PCA przy redukcji do dwóch wymiarów
PC-2
PC-1
80
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
Grupa czerwona
Grupa niebieska
Grupa zielona
Grupa różowa
Grupa pomarańczowa
Centroidy

4. ANALIZA WYNIKÓW I WNIOSKI

4.1 Porównanie klasteryzacji z użyciem PCA i bez

Na podstawie wykresów przedstawiających wyniki klasteryzacji możemy stwierdzić, że grupa czerwona i grupa niebieska prezentują się podobnie. Reszta niestety nie pasuje do siebie, na co mogły mieć wpływ płeć oraz wiek. Można także stwierdzić, że wyniki w przypadku klasteryzacji bez PCA są bardziej skondensowane (grupy występują w konkretnych przedziałach), w przypadku PCA, są one bardziej rozrzucone. Ostatecznie można ustalić, że wyniki klasteryzacji przy użyciu PCA, są trudniejsze do oceny, ze względu na inne wartości, niż te będące w bazie danych oraz brak konkretnych nazw kolumn odwołujących się do tych istniejących w bazie.

4.2 Opis i ocena potencjału każdej z grup

Za klientów docelowych można uznać tych występujących w grupie zielonej, względu na wysoką średnią wartość rocznego dochodu oraz wysoki wskaźnik oceny wydawania, do nich należy kierować najwięcej informacji marketingowych.

Należy także pomyśleć o grupie pomarańczowej, która także posiada wysokie dochody, ale charakteryzuje się niską skłonnością do wydawania swoich pieniędzy, do nich także warto skierować kampanie reklamowe, aby nakłonić ich do rozdysponowania swoich wysokich oszczędności.

Warto wziąć też pod uwagę grupę niebieską, która może nie posiada wysokich zasobów pieniężnych, ale charakteryzuje się wysoką skłonnością do wydawania swoich oszczędności. Do tej grupy warto kierować różnego rodzaju promocje i przeceny, którymi, patrząc na ocenę wydawania i obniżkę cenową, na pewno będą zainteresowani.

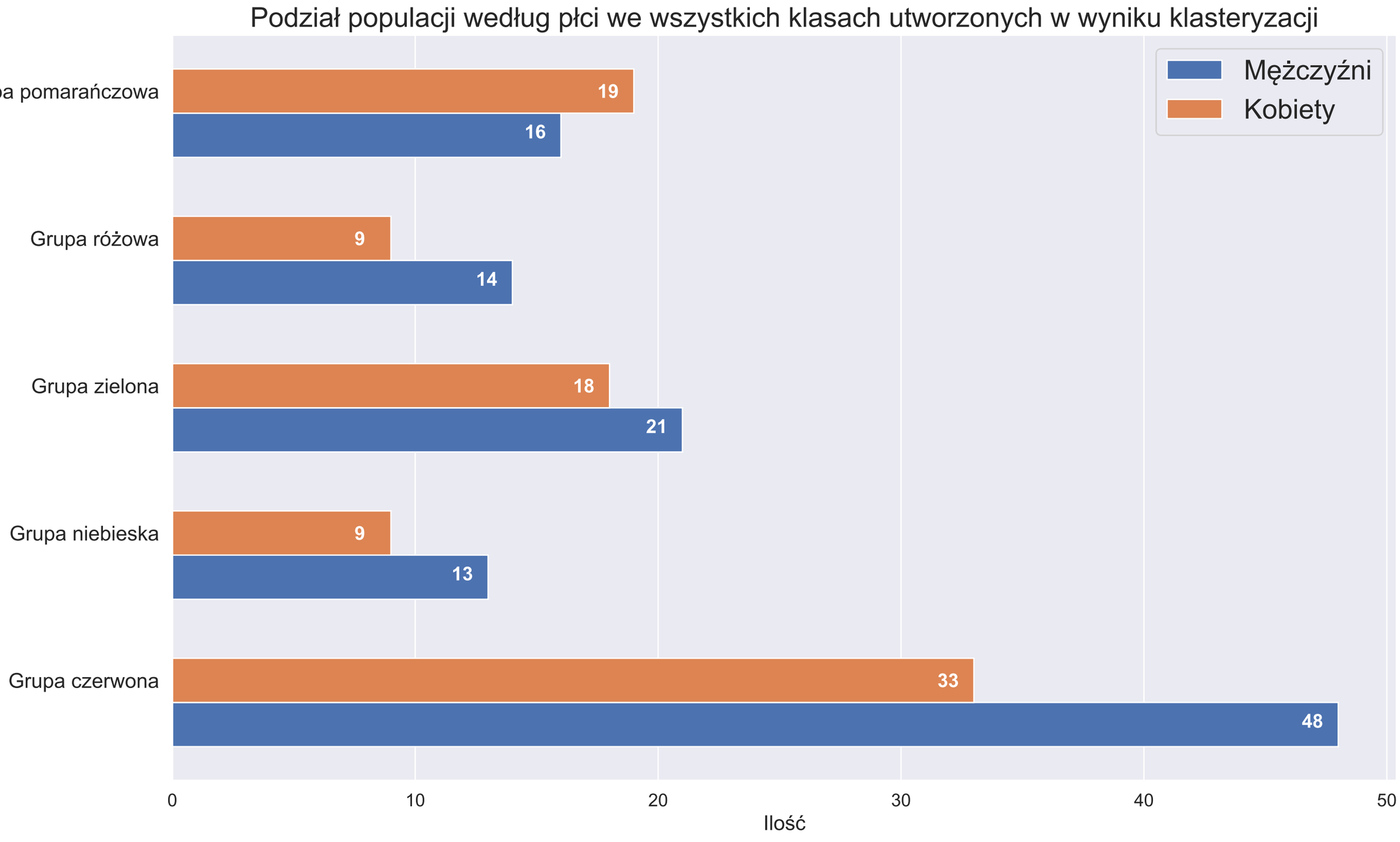
Powinno się także zastanowić nad grupą czerwoną, która jest najbardziej liczebną ze wszystkich grup, posiada średnie zarobki, oraz średnią ocenę wydawania. Te wyniki pozwalają na stały oraz stabilny dochód z tejże grupy, dlatego marketingowo nie należy o nich zapominać.

W grupie różowej powinno się ograniczyć akcje marketingowe i reklamy do niezbędnego minimum. Grupa ta nie może sobie pozwolić na wiele wydatków, a także nie będą skłonni do licznych promocji, ze względu na niski stopień oceny wydawania.

4.3 Kwestia dominacji płci w grupach

Na podstawie poniższego wykresu reprezentującego udział kobiet i mężczyzn w każdej grupie, można stwierdzić, że tylko jedna grupa jest szczególnie zdominowana przez mężczyzn i jest to grupa czerwona (48 do 33). Reszta grup względnie równo zasiedlona przez kobiety jak i mężczyzn.

Gdańsk, 19.05.2021


Podział populacji według płci we wszystkich klasach utworzonych w wyniku klasteryzacji
Mężczyźni
Kobiety
rupa pomarańczowa
19
16
Grupa różowa
9
14
Grupa zielona
18
21
Grupa niebieska
9
13
Grupa czerwona
33
48
0
10
20
30
40
50
Ilość

4.4 Znaczenie oraz wpływ zależności między kolumnami na wynik klasteryzacji

Dzięki licznie wygenerowanym wykresom zależności, można dojść do wielu ciekawych wniosków jak, np.:

- Osoby pomiędzy 30, a 50 rokiem życia posiadają największy roczny dochód
- Wiele osób przed 20 rokiem życia, może pochwalić się rocznymi dochodami na poziomie od 40 do 80 k$
- U osób pomiędzy 60, a 70 rokiem życia rzadko spotykane są roczne dochody na poziomie niższym niż 40 k$
- Najwięcej osób mających mniej niż 20 k$ rocznego dochodu, jest w przedziale od 20 do 25 roku życia
- Najwięcej wydają osoby mające około 30 lat
- Najwięcej osób, które wydają najmniej jest pomiędzy 30, a 60 rokiem życia
- Osoby starsze nie wydają ani za dużo, ani za mało
- Pomiędzy 35, a 50 rokiem życia następuje spadek skłonności do wydawania o około 40 punktów
- Najwięcej jest osób zarabiających pomiędzy 40, a 80 k$ rocznie
- Roczny dochód w wysokości 20 k$ mają osoby wieku 20-30 lat, jak i w wieku 60-70 lat
- Osoby mające roczny dochód na poziomie 0-40 lub 70-140 są najbardziej skłonne do wydawania pieniędzy
- Ocenę wydawania na poziomie 40-60 mają osoby zarabiające od 40 do 70 k$
- Ocenę wydawania od 40 do 60 posiadają osoby w każdym przedziale wiekowym
- Ocenę wydawania od 60 do 100 posiadają tylko osoby w wieku od 20 do 40 lat
- Wykres zależności rocznego dochodu od oceny wydawania wygląda bardzo podobnie do wykresu obrazującego wynik klasteryzacji bez PCA co potwierdza, poprawność wykonanego wykresu

Gdańsk, 19.05.2021