Artur Prasuła, Aleksandra Kowalczyk, Kacper Achramowicz

Podstawy analizy i reprezentacji danych “Grupowanie państw na podstawie wyglądu ich flag”

1. Analiza eksploracyjna zbioru danych

1.1 Poznanie i formatowanie zbioru danych

W celu przygotowania danych do analizy na początku pracy stworzyliśmy ramkę danych korzystając z pliku *flag.data.* Dodane zostały nazwy kolumn, ale też ujednoliciliśmy typ danych w całej ramce, tzn. zamieniliśmy wartości atrybutów kategorycznych na liczby dziesiętne.

name	landmass	zone	area	population	language	religion	bars	stripes	colours	red	...	saltires	quarters	sunstars	crescent	triangle	icon	animate	text	t
Afghanistan	5	1	648	16	10	2	0	3	5	1	...	0	0	1	0	0	1	0	0	
Albania	3	1	29	3	6	6	0	0	3	1	...	0	0	1	0	0	0	1	0	
Algeria	4	1	2388	20	8	2	2	0	3	1	...	0	0	1	1	0	0	0	0	
American-Samoa	6	3	0	0	1	1	0	0	5	1	...	0	0	0	0	1	1	1	0	
Andorra	3	1	0	0	6	0	3	0	3	1	...	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Western-Samoa	6	3	3	0	1	1	0	0	3	1	...	0	1	5	0	0	0	0	0	
Yugoslavia	3	1	256	22	6	6	0	3	4	1	...	0	0	1	0	0	0	0	0	
Zaire	4	2	905	28	10	5	0	0	4	1	...	0	0	0	0	0	1	1	0	

Zrzut 1 Ramka danych zawierających zbiór z pliku

Następnym krokiem było sprawdzenie czy dane posiadają jakieś braki. Wnioskiem z tego testu było, że zbiór jest pełny. Wynik pokrył się z umieszczonym na stronie https://archive.ics.uci.edu/ml/datasets/Flags.

```
landmass      False          orange        False
zone          False          mainhue       False
area          False          circles       False
population    False          crosses       False
language      False          saltires      False
religion      False          quarters      False
bars          False          sunstars      False
stripes       False          crescent      False
colours       False          triangle      False
red           False          icon          False
green         False          animate       False
blue          False          text          False
gold          False          topleft       False
white         False          botright      False
black         False          dtype: bool
```

Zrzut 2 Test sprawdzający, czy występują puste dane. False oznacza, że dane są kompletne.

1.2 Analiza pojedynczych atrybutów

Dokonaliśmy również analizy pojedynczych atrybutów, ponieważ mogło okazać się to przydatne w dalszej analizie zbioru, głównie przy wyciąganiu wniosków z histogramu. Dzięki temu łatwiej było stwierdzić najczęściej występującą ilość obiektów poszczególnych atrybutów (np. najbardziej popularną ilość kolorów we fladze).

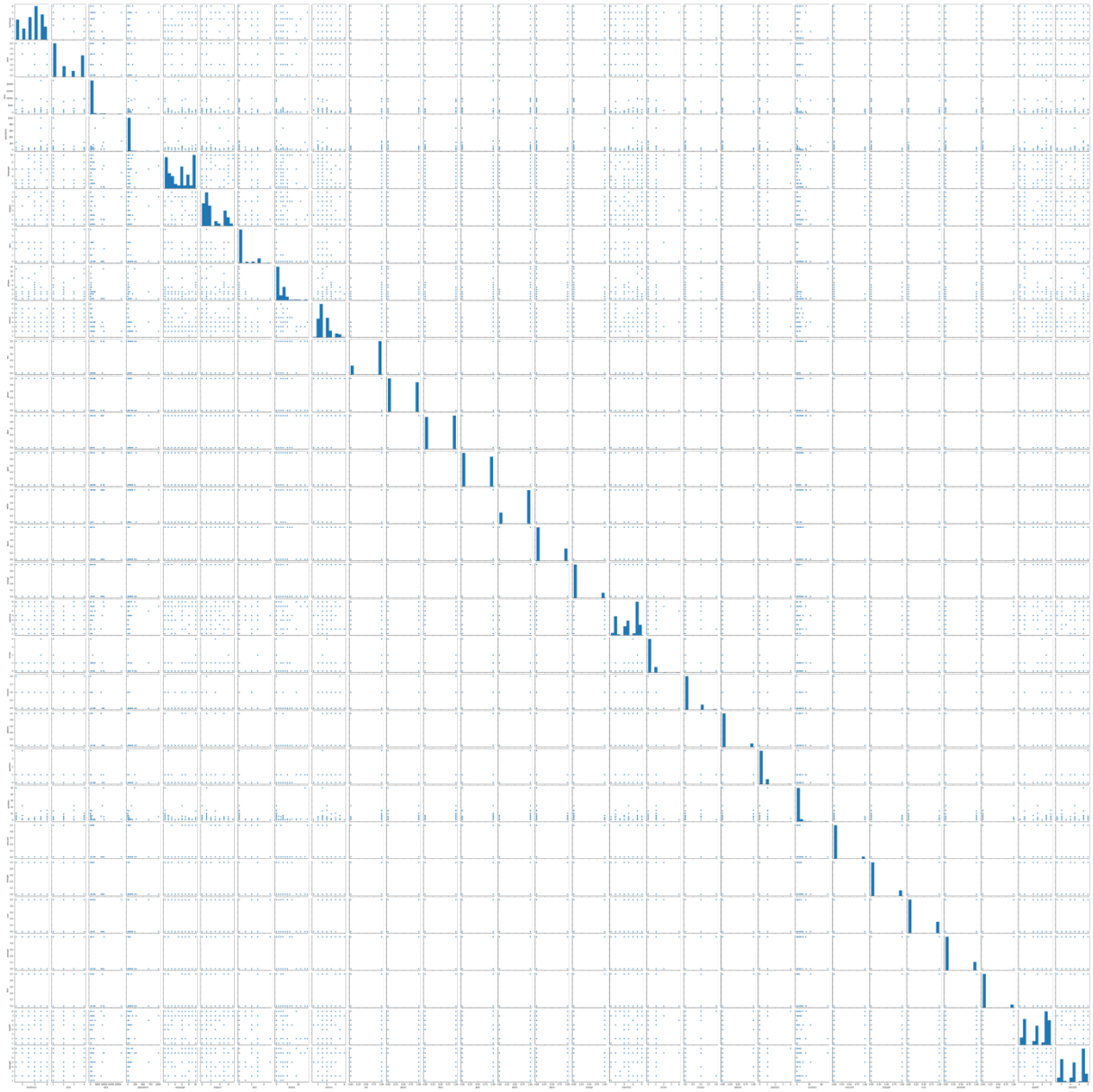
	landmass	zone	area	population	language	religion	bars	stripes	colours	red	...	circles	crosse
count	194.000000	194.000000	194.000000	194.000000	194.000000	194.000000	194.000000	194.000000	194.000000	194.000000	...	194.000000	194.00000
mean	3.572165	2.211340	700.046392	23.268041	5.340206	2.190722	0.453608	1.551546	3.463918	0.788660	...	0.170103	0.14948
std	1.553018	1.308274	2170.927932	91.934085	3.496517	2.061167	1.038339	2.328005	1.300154	0.409315	...	0.463075	0.38538
min	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	...	0.000000	0.00000
25%	3.000000	1.000000	9.000000	0.000000	2.000000	1.000000	0.000000	0.000000	3.000000	1.000000	...	0.000000	0.00000
50%	4.000000	2.000000	111.000000	4.000000	6.000000	1.000000	0.000000	0.000000	3.000000	1.000000	...	0.000000	0.00000
75%	5.000000	4.000000	471.250000	14.000000	9.000000	4.000000	0.000000	3.000000	4.000000	1.000000	...	0.000000	0.00000
max	6.000000	4.000000	22402.000000	1008.000000	10.000000	7.000000	5.000000	14.000000	8.000000	1.000000	...	4.000000	2.00000

```
1 flags.mode()
```

	landmass	zone	area	population	language	religion	bars	stripes	colours	red	...	saltires	quarters	sunstars	crescent	triangle	icon	animate	text
0	4	1	0	0	10	1	0	0	3	1	...	0	0	0	0	0	0	0	0

Zrzut 3 Tabela miar pojedynczych atrybutów

Następnie utworzyliśmy macierz wykresów punktowych i histogramy dla każdego atrybutu.



Zrzut 4 Macierz wykresów punktowych i histogramy atrybutów

Analiza histogramów i tablicy miar pojedynczych atrybutów pozwoliła nam sformułować pewne wnioski, najciekawsze z nich to:

- najwięcej krajów znajduje się na kontynencie Afryki, następnie Azji, a potem Europy
- histogramy pokazują, że zdecydowana większość krajów, ma powierzchnię mniejszą niż 2500 tysięcy kilometrów
- histogram populacji pokazuje, że zdecydowana większość krajów ma populacje mniejszą niż 125 milionów
- histogram języków pokazuje, że najpopularniejszym językiem jest język angielski
- w większości krajów językiem urzędowym jest język z grupy określonej jako ‘inne’
- w większości krajów dominującym wyznaniem jest odłam chrześcijaństwa niebędący katolicyzmem
- kolorem najczęściej występującym w prawym górnym oraz lewym dolnym rogu jest czerwony
- z histogramu występowania koloru czerwonego możemy wnioskować, że ponad ¾ flag zawiera kolor czerwony
- z histogramu występować kolory zielonego, możemy wywnioskować, że prawie połowa flag zawiera ten kolor
- z histogramu występować kolory niebieskiego, możemy wywnioskować, że ponad połowa flag zawiera ten kolor
- z histogramu występować kolory złotego (też żółtego), możemy wywnioskować, że trochę prawie połowa flag zawiera ten kolor
- z histogramu występowania koloru białego możemy wnioskować, że około ¾ flag zawiera kolor biały
- z histogramu występowania koloru czarnego możemy wnioskować, że około ¾ flag nie zawiera koloru czarnego
- z histogramu występowania koloru pomarańczowego (też brązowego) możemy wnioskować, że niewielka ilość flag zawiera ten kolor
- histogram części pionowych i poziomych pokazuje, że większość flag nie zawiera żadnych pionowych i poziomych słupków
- z histogramu atrybutu ilości kolorów we fladze: najczęściej występują flagi z 3 kolorami
- z histogramu atrybutu: koła, krzyże, krzyże przekątne, ćwiartki, słońce-gwiazdy, można wywnioskować, że zdecydowana większość flag ich nie zawiera
- analizując histogramy atrybutów 24-28 możemy wywnioskować, że zdecydowana większość flag nie zawiera tych znaków, jeśli jednak weźmiemy pod uwagę flagi je zawierające, możemy przedstawić następujące relacje przedstawiające częstotliwość występowania tych atrybutów: ikony > animacje > trójkątny > tekst > półksiężyc.

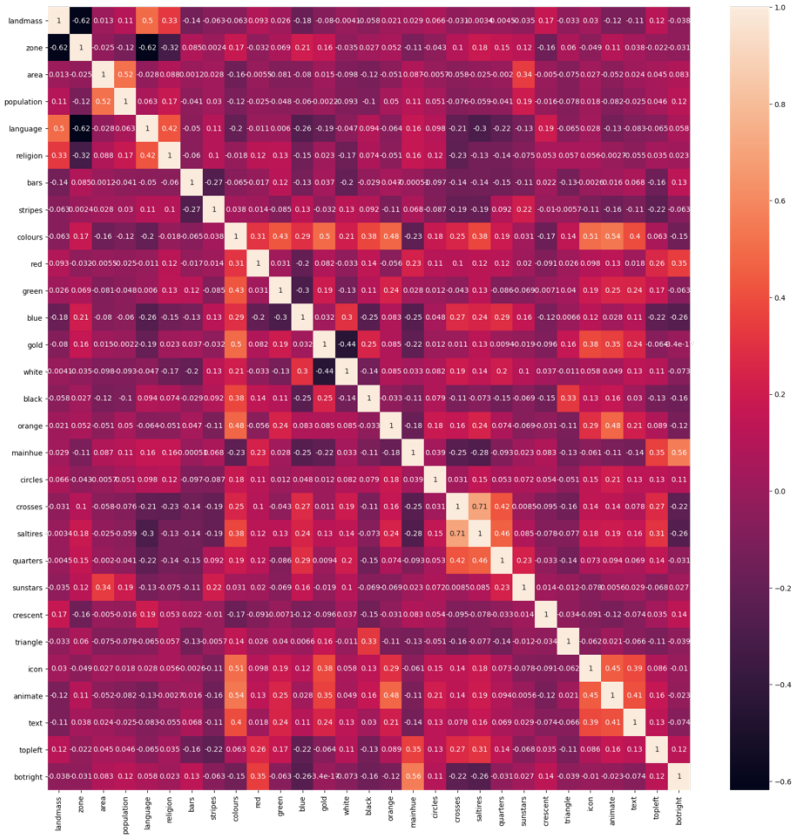
1.3 Analiza wykresów macierzy punktowych

Analiza wykresów punktowych doprowadziła nas do ciekawych spostrzeżeń. Najbardziej interesujące to:

- wykres punktowy słońce-gwiazda i kontynentów zawiera wartość szczególnie odróżniającą się od innych- wartość nietypowa. Jest to flaga z 50 gwiazdami na kontynencie Ameryki Północnej. Wiemy więc, że ten obiekt reprezentuje flagę USA.
- znak półksiężyca występuje jedynie na flagach z kontynentu Afryki i Azji
- wszystkie kraje z religią marksistowską mają kolor czerwony
- wszystkie kraje z religią hinduizmu i 'inną' nie mają tekstu ani animacji we fladze
- znak półksiężyca we fladze mają kraje z religią muzułmańską, buddyjską, hinduską i marksistowską
- z wykresów punktowych atrybutów, z których jeden to ilość kolorów występujących we fladze wnioskujemy, że wszystkie flagi, które mają największą ilość kolorów zawierają kolor czerwony, niebieski, złoty i biały.

1.4 Analiza korelacji pomiędzy atrybutami

Kolejnym krokiem analizy eksploracyjnej było utworzenie macierzy korelacji dla wszystkich atrybutów.



Zrzut 5 Macierz korelacji atrybutów w postaci mapy ciepła

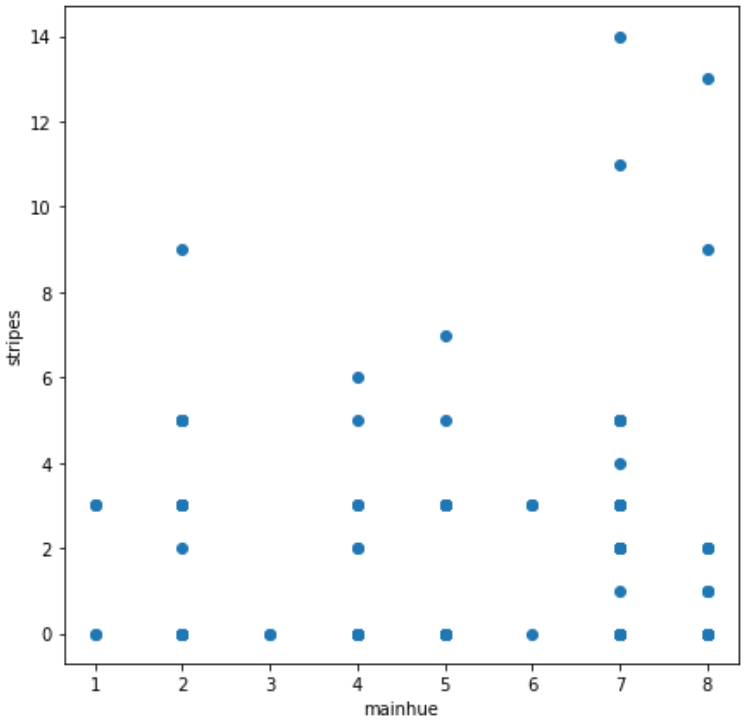
Możemy zauważyć pewne powiązania między poszczególnymi atrybutami:

- korelacja wysoka występuje między kontynentem a strefą i strefą a językiem
- korelacja umiarkowana występuje między kontynentem a językiem, między powierzchnią i liczebnością populacji a także między religią a językiem
- korelacja niska między kontynentem a religią, a także między strefą a religią
- powyższe korelacje doprowadzają nas do wniosku, że język jest bardziej powiązany z kontynentem i język z religią niż religia z kontynentem.
- Istnieją umiarkowane korelacje między atrybutem mówiącym o ilości kolorów we fladze a atrybutem ikony, animacje, tekst
- powyższa obserwacja doprowadza nas do wniosku, że obecność ikon/animacji/tekstu zwiększa prawdopodobieństwa, że flaga składa się z wielu kolorów
- najwyższa korelacja umiarkowana między atrybutem mówiącym o ilości kolorów we fladze a kolorem złotym– im więcej kolorów w fladze, tym większe prawdopodobieństwo, że wystąpi tam właśnie kolor złoty (żółty)
- następnie korelacja umiarkowana między atrybutem mówiącym o ilości kolorów we fladze a kolorem pomarańczowym– im więcej kolorów w fladze, tym większe prawdopodobieństwo, że wystąpi tam właśnie pomarańczowy
- korelacja umiarkowana między atrybutem mówiącym o ilości kolorów we fladze a kolorem zielonym– im więcej kolorów w fladze, tym większe prawdopodobieństwo, że wystąpi tam właśnie zielony
- jeśli zaś chodzi o kolory: czerwony, niebieski, biały, to nie ma tutaj aż takiego silnego związku ze zwiększeniem ilości kolorów we fladze- warto zauważyć, że te 3 kolory to właśnie 3 najczęściej występujące kolory we flagach- co uzasadnia, dlaczego nie istnieje tu silna korelacja między ich występowaniem a atrybutem mówiącym o ilości kolorów we fladze
- korelacja niska między zielonym a niebieskim, czyli im większa częstotliwość występowania jednego koloru, tym mniejsze prawdopodobieństwo wystąpienia drugiego
- korelacja niska między niebieskim a białym, czyli im większa częstotliwość występowania jednego koloru, tym większe prawdopodobieństwo wystąpienia drugiego
- korelacja umiarkowana między złotym a białym, czyli im większa częstotliwość występowania jednego koloru, tym mniejsze prawdopodobieństwo wystąpienia drugiego
- korelacja niska między czarnym a trójkątami, czyli możemy dostrzec związek między występowaniem koloru czarnego, a występowaniem trójkątów we fladze

- korelacja umiarkowana między pomarańczowym a animacjami, czyli możemy dostrzec związek między występowaniem koloru pomarańczowego (i brązowego), a występowaniem animacji we fladze
- korelacja wysoka między krzyżami pionowymi a przekątnymi – im więcej krzyży pionowych na fladze, tym większe prawdopodobieństwo, że na fladze znajduje się także krzyże przekątne
- korelacja umiarkowana między krzyżami a ćwiartkami – im więcej krzyży na fladze, tym większe prawdopodobieństwo, że flaga podzielona jest na ćwiartki
- korelacja umiarkowana między półksiężycami a ćwiartkami – im więcej półksiężycy na fladze, tym większe prawdopodobieństwo, że flaga podzielona jest na ćwiartki
- korelacja umiarkowana między ikonami a animacjami – im więcej ikon na fladze, tym większe prawdopodobieństwo, że na fladze znajduje się także animacji
- korelacja niska między ikonami a tekstem – im więcej ikon na fladze, tym większe prawdopodobieństwo, że na fladze znajduje się także tekst
- korelacja umiarkowana między animacjami a tekstem – im więcej animacji na fladze, tym większe prawdopodobieństwo, że na fladze znajduje się także tekst.

1.5 Wybór atrybutów do dalszych działań

Analizując histogramy poszczególnych atrybutów zdecydowaliśmy, że jednym z atrybutów, które wybierzemy do dalszej pracy, będzie mainhue (mówiący o głównym kolorze flagi). Przyjmuje on 8 różnych wartości, każdy o innej częstotliwości występowania. Przy wyborze drugiego atrybutu rozważaliśmy colours (mówi o ilości kolorów występujących we fladze) i stripes (mówi o ilości poziomych bloków we fladze). Nie braliśmy pod uwagę atrybutu bars (mówiącego o ilości pionowych bloków we fladze), ponieważ jego wartości były mniej różnorodne i bardziej skupione niż dla atrybutu stripes. Atrybut colours, mimo że jego wartości były bardziej rozproszone został przez nas odrzucony. Doszliśmy do wniosku, że w połączeniu z mainhue dla części grup nie dostalibyśmy wiarygodnych wyników- wizualnie grupy nie byłyby do siebie podobne, ponieważ ilość kolorów wszelkiego rodzaju symboli wpływa na wielkość atrybutu, lecz w o wiele mniejszym stopniu na wygląd flagi, co zaburzałaby spójność grup. Atrybut stripes ma różnorodne wartości, a także posiada realny wpływ na wygląd flagi- dlatego zdecydowaliśmy się właśnie na niego.



Zrzut 6 Zależność wybranych przez nas atrybutów

2. Grupowanie danych

Na podstawie wybranych atrybutów w analizie eksploracyjnej, przeprowadziliśmy grupowanie danych przy użyciu dwóch metod:

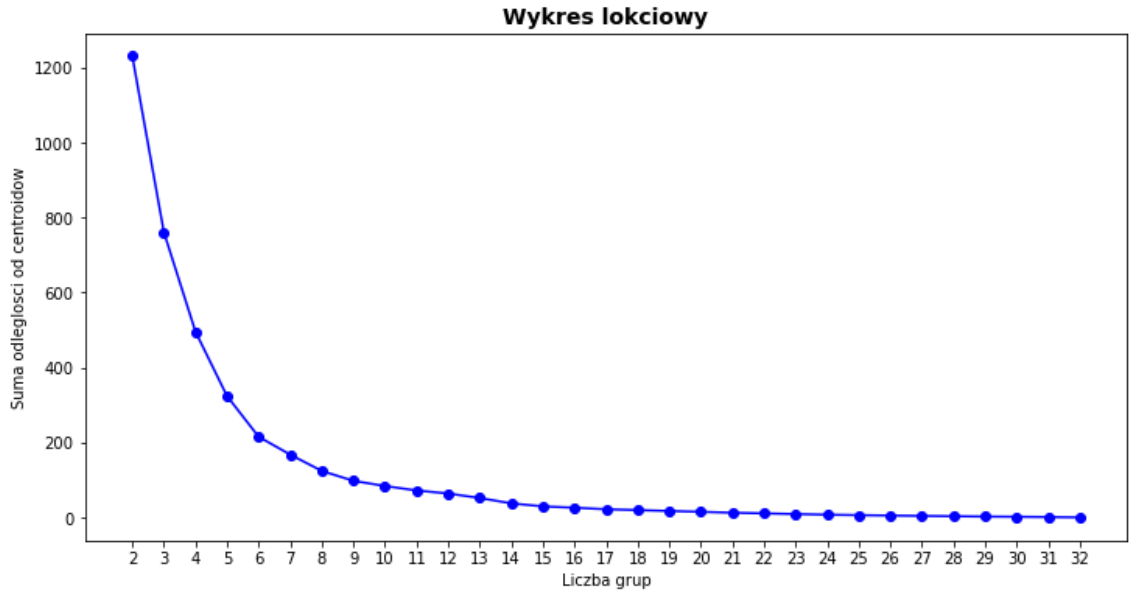
- Metody k-średnich
- Metody hierarchicznego grupowania aglomeracyjnego

Dzięki pogrupowaniu za pomocą dwóch metod, możemy porównać wyniki z obu metod i wybrać te wyniki, które lepiej spełniają nasze oczekiwania.

2.1. Metoda k-średnich

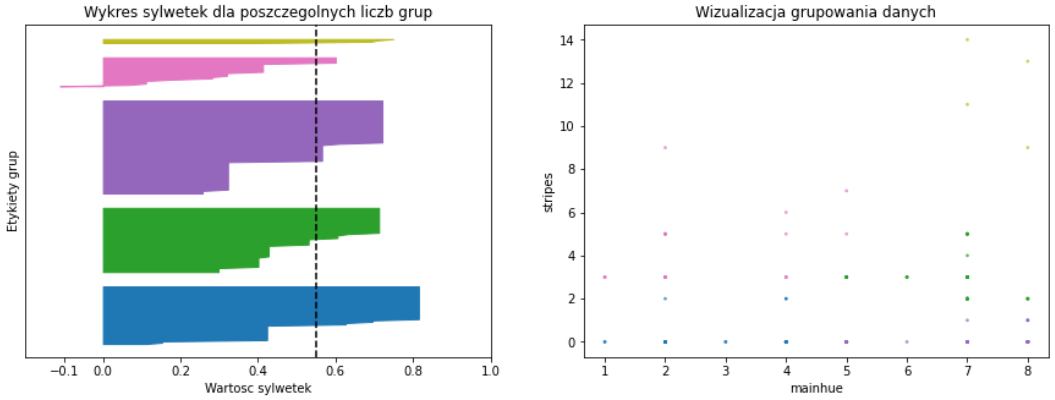
Pierwszą metodą, którą użyliśmy, była metoda k-średnich, która grupuje dane na podstawie odległości punktów od centroidów. Wybierając tę metodę, mieliśmy obawy czy poradzi sobie ona dobrze dla atrybutu kategorycznego takiego jak mainhue. Mimo obaw postanowiliśmy wykonać przy jej pomocy grupowanie, ponieważ atrybut stripes jest atrybutem ilościowym.

Metoda ta charakteryzuje się tym, że ilość grup ustawiana jest odgórnie. Dlatego wykonaliśmy analizę dla liczby grup z przedziału 2-32 (32 – liczba kombinacji wybranych atrybutów). Odpowiednią, wynikową liczbę grup wybraliśmy analizując wykres sylwetek dla poszczególnych grup oraz wykres łokciowy.



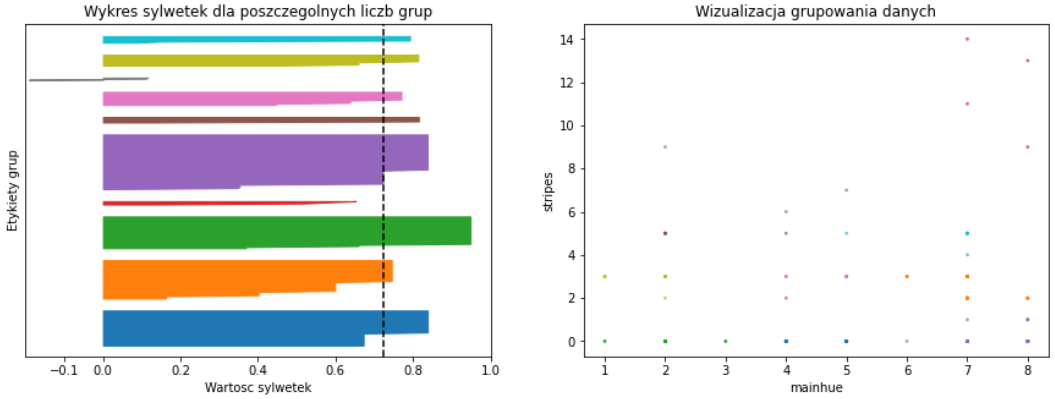
Zrzut 7 Wykres łokciowy dla metody k-średnich

Analiza sylwetki k = 5



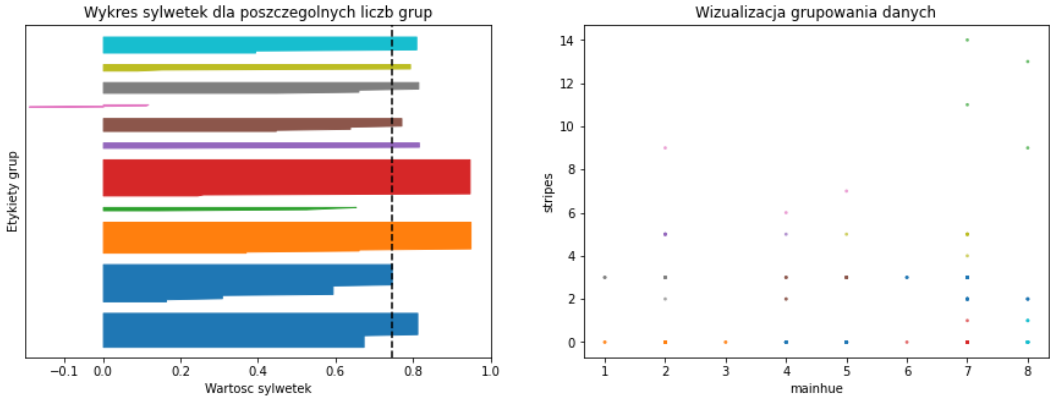
Dla n = 5, średnia wartosc sylwetki wynosi: 0.551, Suma odleglosci od centroidow: 323.56

Analiza sylwetki k = 10



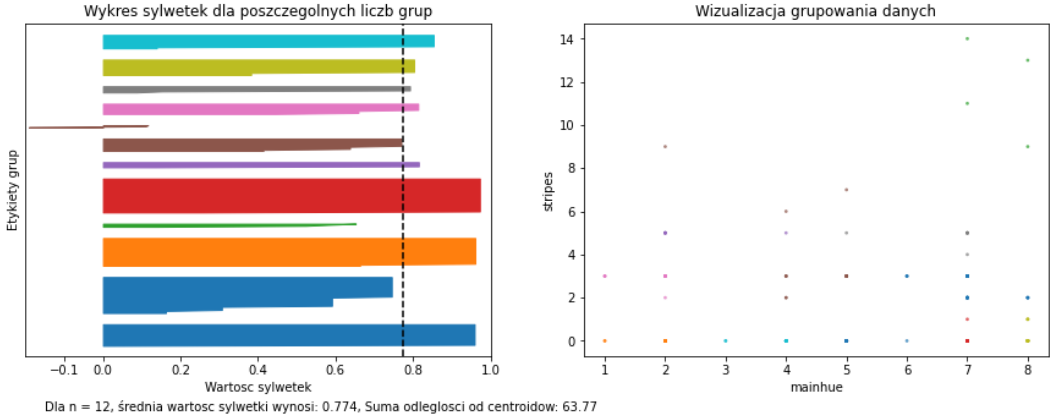
Dla n = 10, średnia wartosc sylwetki wynosi: 0.724, Suma odleglosci od centroidow: 83.75

Analiza sylwetki k = 11



Dla n = 11, średnia wartosc sylwetki wynosi: 0.745, Suma odleglosci od centroidow: 71.86

Analiza sylwetki k = 12

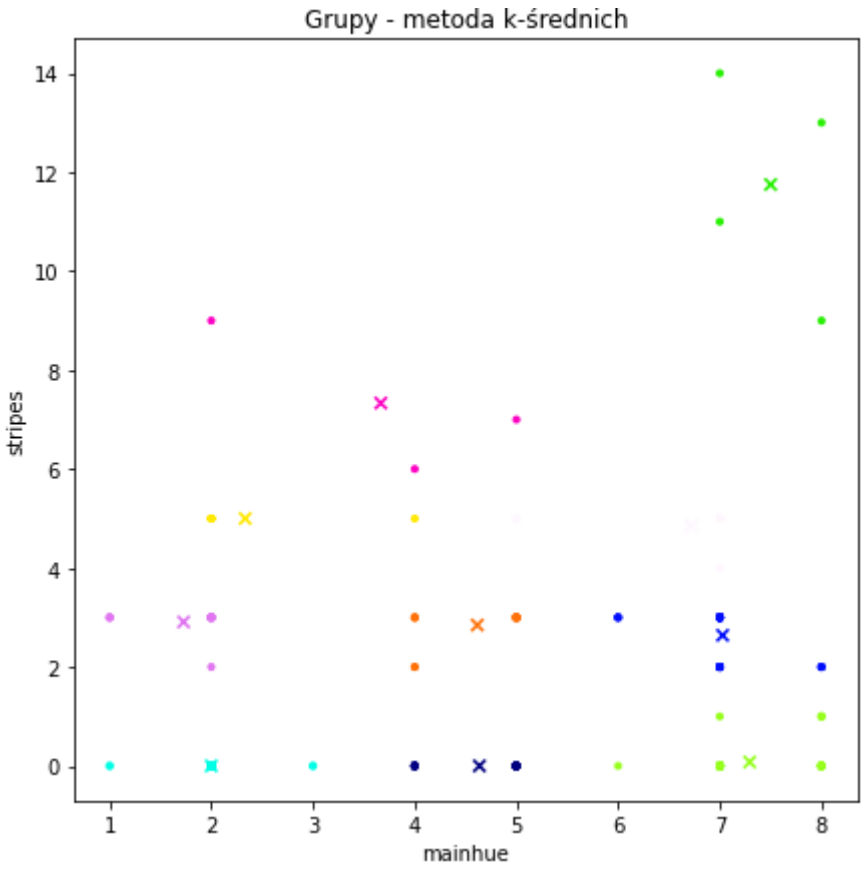


Dla n = 12, średnia wartosc sylwetki wynosi: 0.774, Suma odleglosci od centroidow: 63.77

Wnioski o ilości grup:

- Wykres łokciowy łagodnie zwalnia, dlatego ciężko znaleźć optymalną liczbę grup.
- Dla ilości grup 6-9 widać, że „spadek” wykresu maleje.
- Z wykresu łokciowego wynika, że optymalna ilość grup mieści się w przedziale 10-12 (włącznie).
- Z analizy sylwetki wynika, że optymalna ilość grup jest równa 5.
- Wykres sylwetek dla liczby grup w przedziale 10-12 (włącznie) nie jest idealny, ale akceptowalny.
- Wykresy sylwetek dla k bliskiego lub równego 32 są najlepsze, ale spowodowane jest to tym, że 32 jest to liczba możliwych kombinacji wybranych atrybutów (punktów na wykresie).
- Z obu tych analiz możemy stwierdzić, że optymalna ilość grup jest równa 10.

Wynik grupowania (liczba grup – 10):



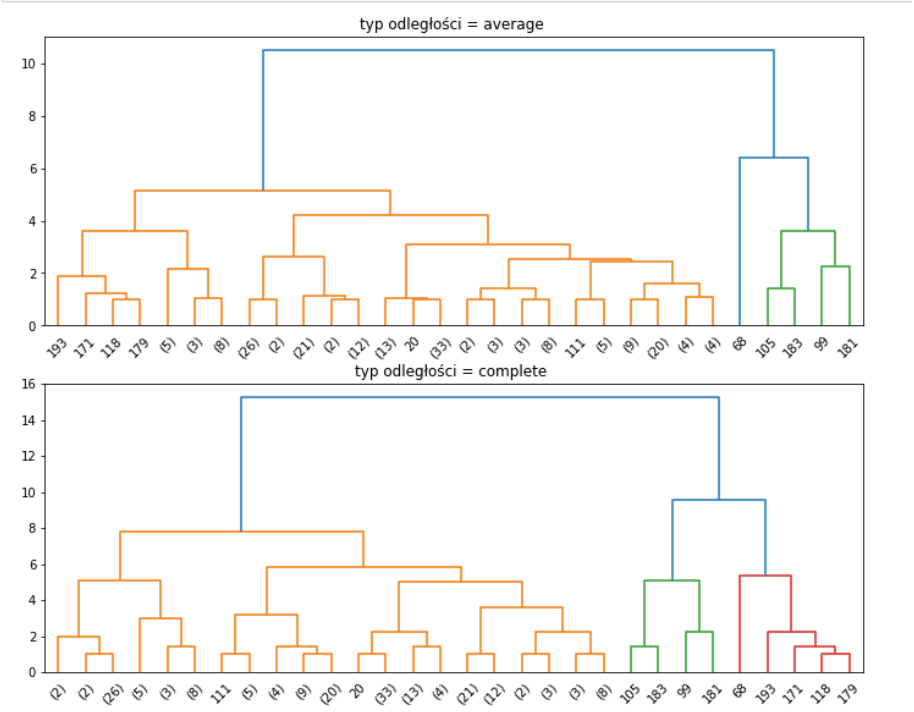
Zrzut 8 Wynik grupowania dla liczby grup równej 10

Grupowanie danych przebiegło pomyślnie. Zostały stworzone grupy, których punkty są blisko siebie, co teoretycznie gwarantuje nam podobieństwo w wyglądzie flag w grupach.

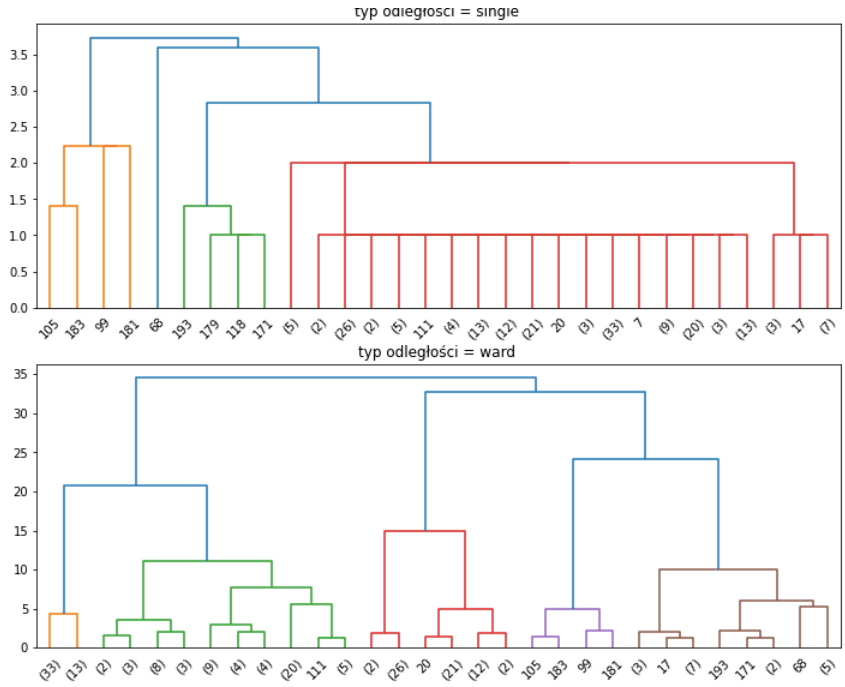
Możemy zauważyć, że nie wszystkie punkty w jednej grupie mają tą samą wartość atrybutu mainhue, co nie jest wskazane (ale akceptowalne), ponieważ to kolor flagi jest tą cechą, którą najłatwiej jest dostrzec.

2.2 Metoda hierarchicznego grupowania aglomeracyjnego

Pracę nad hierarchicznym grupowaniem aglomeracyjnym rozpoczęliśmy od analizy i wyboru sposobu określenia odległości między skupiskami obiektów. W tym celu stworzyliśmy 4 dendrogramy, każdy dla jednego typu odległości.



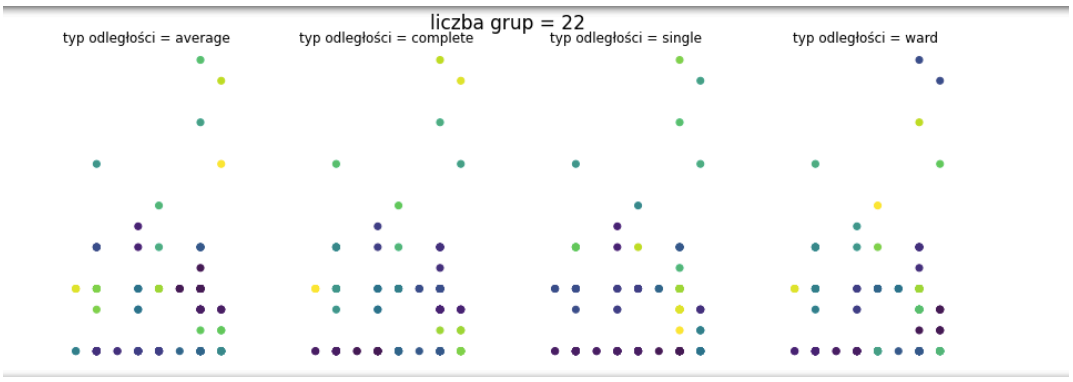
Zrzut 9 Przykładowe dendrogramy dla typu odległości równej average i complete.



Zrzut 10 Przykładowe dendrogramy dla typów odległosci równych single i ward.

Na początku wykluczyliśmy typ single, ponieważ bardzo mocno uśredniał wyniki, połączył większość różnych skupisk obiektów w jedną grupę. Spowodowało to, że wynikiem grupowania było kilka bardzo licznych grup, gdzie dana grupa zawierała obiekty znacząco różniące się od siebie- ten sposób nie odpowiadałby efektom, jakie chcieliśmy uzyskać. Następnie odrzuciliśmy typ average, ponieważ wiele różnych obiektów w tej metodzie zostało potraktowanych jako podobne i połączone w jedną grupę, przez co otrzymaliśmy bardzo uśrednione wyniki. Typ ward i complete dawała bardzo podobny wynik – podobną liczbę grup i wygląd dendrogramu. Analizując jednak grupy doszliśmy do wniosku, że lepsze przyporządkowanie daje nam typ complete. tworzy grupy obiektów faktycznie podobnych do siebie.

Następnie stworzyliśmy wykresy pokazujące rozmieszczenie skupisk obiektów i przynależność do grup dla różnej liczebności grup i różnych typów odległości.

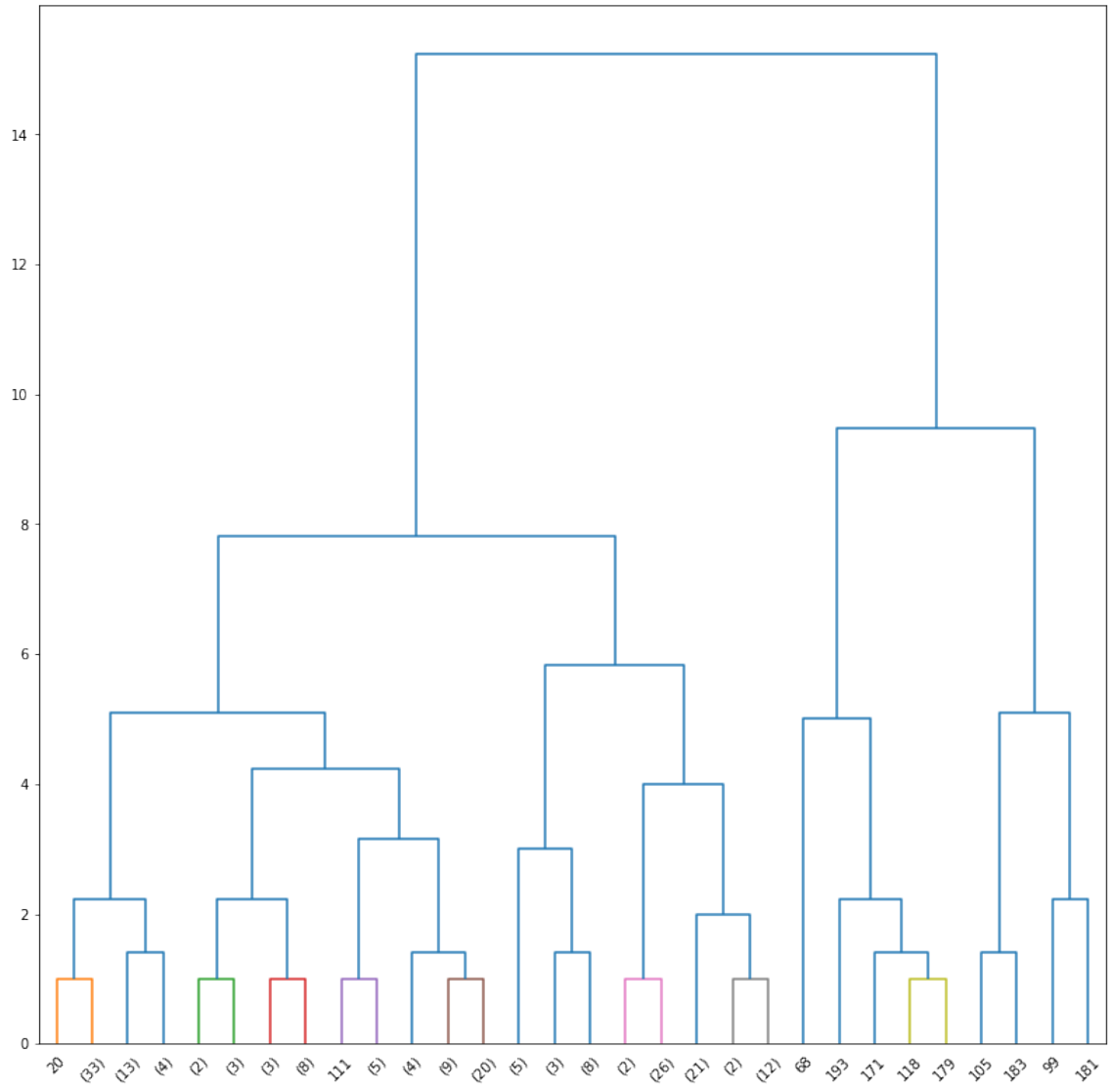


Zrzut 11 Wykres pokazujący skupiska obiektów dla różnych typów odległości dla ilości grup równej 22

Z wykresów odczytaliśmy również informację, że ilość kombinacji naszych dwóch atrybutów (mainhue- główny kolor we fladze i stripes- ilość poziomych części flagi) wynosi 32.

Wykresy potwierdziły naszą decyzje o wykluczeniu metody single- widzimy, że tworzy przede wszystkim dwie grupy o bardzo dużej ilości różnych obiektów. Między wykresami dla average, complete i ward istnieją drobne różnice, polegające na tym, że pewne obiekty są przypisywane do innych grup. Widzimy jednak, że większość obiektów z drobnymi wyjątkami tworzy dla każdego sposobu podobne grupy. Analizując wykresy dla wybranego przez nas typu complete widzimy, że grupy tworzone są z obiektów, które faktycznie znajdują się blisko siebie. Nie mamy tutaj obiektów, które zostały przypisane do jakiejś grupy, mimo tego, że nie znajdują się w jej sąsiedztwie.

Ilość 22 grup wydaje nam się optymalną wielkością, ponieważ bardziej odległe elementy tworzą oddzielne grupy, jednak wciąż część skupisk, które znajdują się blisko siebie tworzą jedną grupę. Mniejsza ilość grup dawała nam zbyt uśrednione wyniki- przez co flagi, które znajdowały się w jednej grupie nie były wizualnie do siebie podobne.



Zrzut 12 Dendrogram dla wybranego typu complete i wartości progu równej 7

Dendrogram pokazuje dokładnie, jak połączone zostały skupiska obiektów, dla wybranego przez nas progu równego 7. Pozioma oś dendrogramu daje nam informacje o liczebności poszczególnych skupisk obiektów. Analizując dendrogram należy zwrócić uwagę, że zostało utworzone 8 grup, z których każda składa się z 2 skupisk obiektów. 7 grup jest odzwierciedleniem poszczególnego skupiska obiektów. Pozostałe 7 grup składa się każda z pojedynczego obiektu, których cechy były na tyle różne od pozostałych, że flagi te, nie mogły zostać zakwalifikowane do żadnej z reszty grup i utworzyły indywidualne grupy.

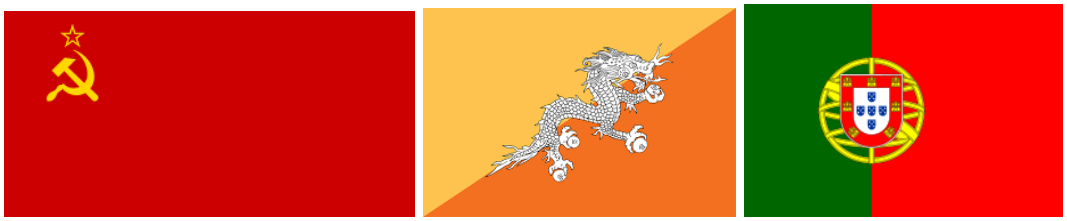
2.3 Porównanie obu metod grupowania

Po analizie wyników doszliśmy do wniosku, że pierwszy sposób grupowania jest mniej efektywny, ponieważ niektóre flagi z jednej grupy są do siebie mało podobne. Z metody k-średnich wynika, że powinniśmy utworzyć od 10 do 12 grup, natomiast metoda hierarchicznego grupowania aglomeracyjnego pozwoliła nam porównać wynik grupowania dla 17 i 22 grup (mniejsza ilość grup została wykluczona). W tym wypadku większa ilość grup drugiej metody pozwala na większą dokładność.

Przykładowe flagi z grupy nr 1 w metodzie k-średnich:

Flaga 1: Andora *Flaga 2: Brazylia* *Flaga 3: Bangladesz*

Przykładowe flagi z grupy nr 1 w hierarchicznym grupowaniu:



Flaga 4: ZSRR *Flaga 5: Bhutan* *Flaga 6: Portugalia*

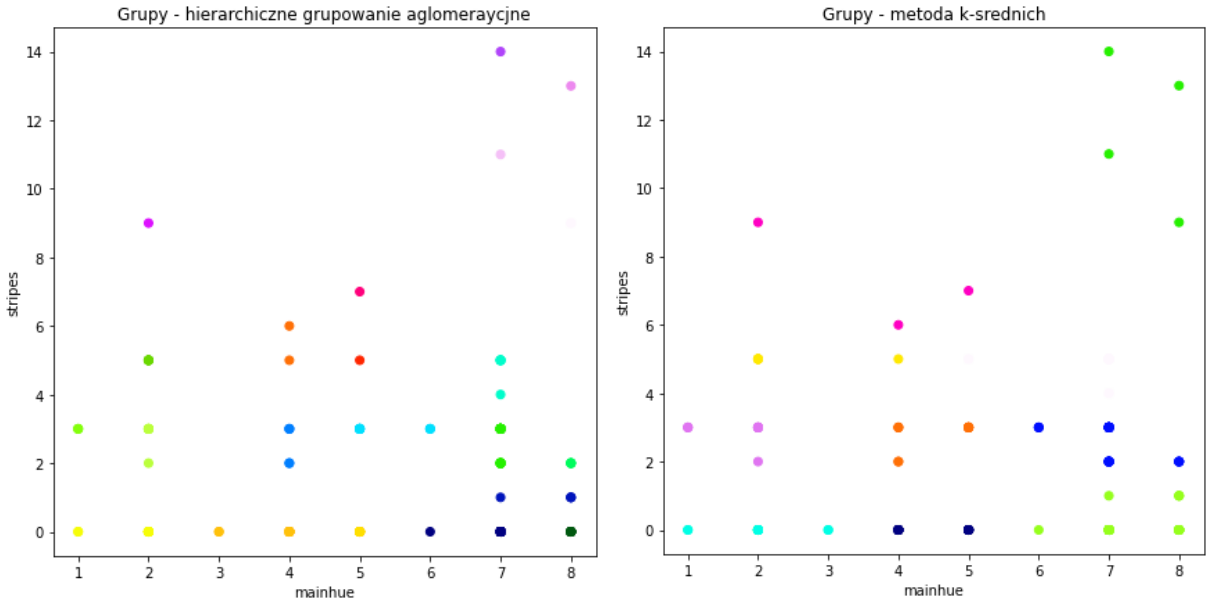
Przykładowe flagi z grupy nr 10 w metodzie k-średnich:

Flaga 7: Gambia *Flaga 8: Puerto Rico* *Flaga 9: Surinam*

Przykładowe flagi z grupy nr 10 w hierarchicznym grupowaniu:

Flaga 10: RFN *Flaga 11: Jordania* *Flaga 12: Egipt*

Jak widać w metoda k-średnich początkową grupę tworzy porównywalnie jakościowo do metody hierarchicznego grupowania aglomeracyjnego, podobnie jest w grupie numer 10. Większość flag jest grupowana prawidłowo.



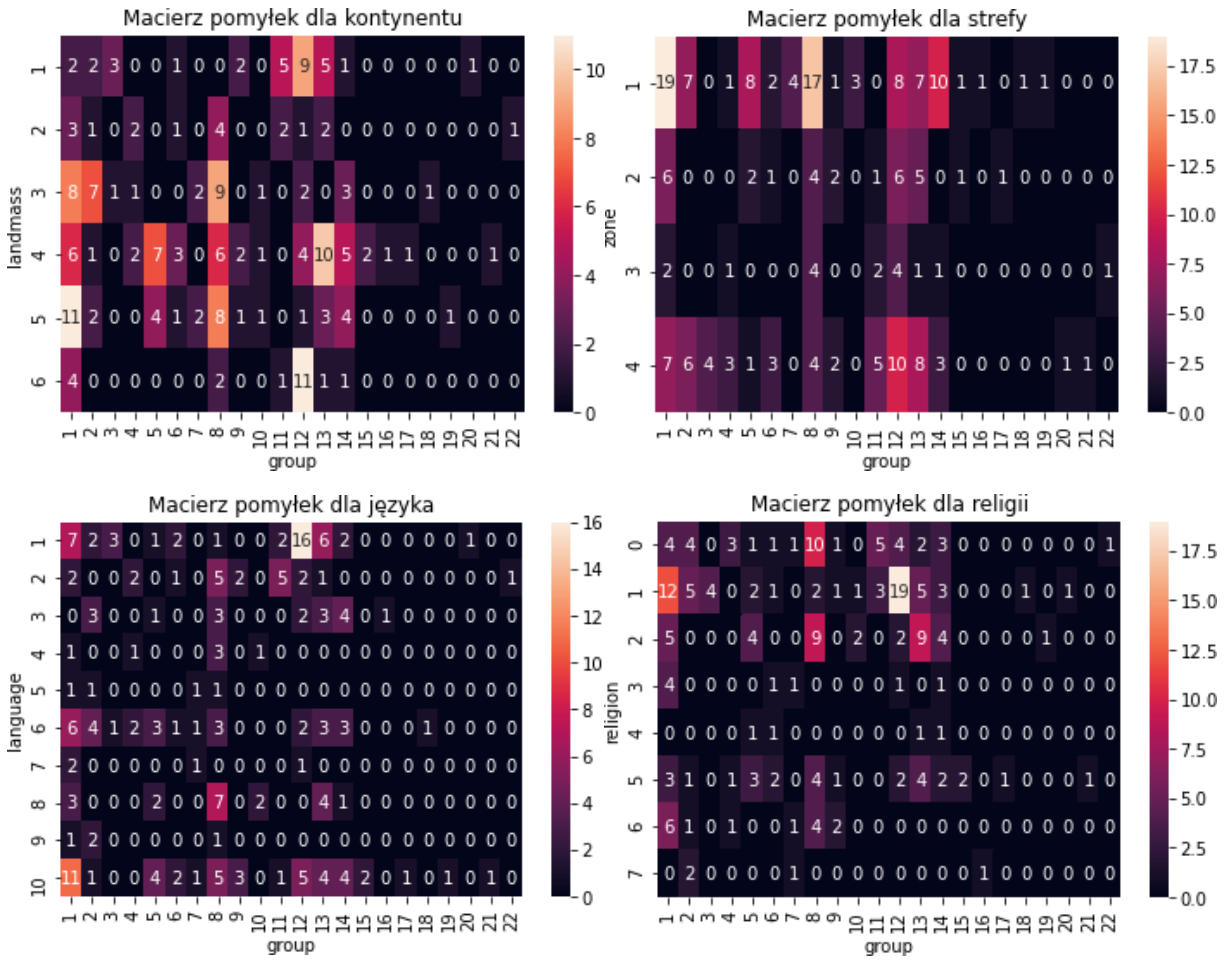
Wykresy grup dla obu metod pokazują, że metoda k-średnich umieszcza częściej w grupach flagi mające różne ilości poziomych pasków jak i mające inny kolor przewodni. Na tej podstawie można stwierdzić, że to kolejny argument pokazujący wyższość metody grupowania aglomeracyjnego w rozpatrywanym przez nas problemie.

3 Wnioski z pogrupowania

3.1 Wnioski z macierzy pomyłek

Po wybraniu odpowiedniej metody i liczby grup, dokonaliśmy analizy wyników. Pierwszym etapem, była analiza za pomocą macierzy pomyłek. Użyliśmy tego by sprawdzić czy istnieją powiązania w wyglądzie flag z:

- położeniem geograficznym,
- panującą w państwie główną religią,
- panującym w państwie głównym językiem.



Wnioski wyciągnięte z macierzy pomyłek znajdują się w „Wizualizacja i analiza efektów grupowania” (niżej).

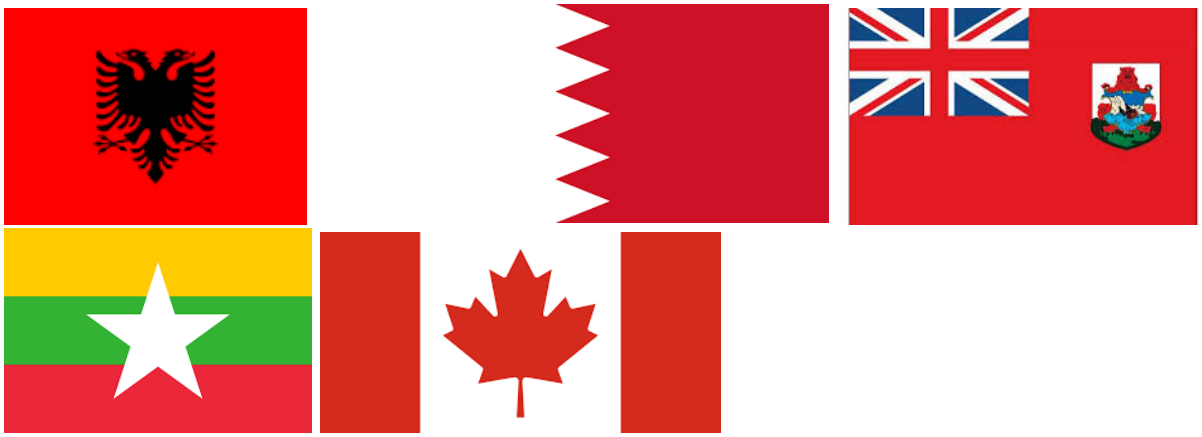
Kolejnym etapem analizy wyników, była analiza pojedynczych atrybutów w zależności od grup. Dokonaliśmy tego, aby sprawdzić czy istnieją powiązania pomiędzy wartościami atrybutów, a grupami oraz sprawdzić jakie wartości przyjmują atrybuty mainhue i stripes (wybrane przez nas atrybuty), w grupach wynikowych.

Ostatnim etapem analizy efektów grupowania była analiza ręczna, która polegała na wybraniu przez nas losowo kilku flag z każdej z grup i sprawdzeniu jak flagi wizualnie do siebie pasują.

3.2 Wizualizacja i analiza efektów grupowania:

Grupa nr 1:

Przykładowe flagi z tej grupy:



Składa się z 34 flag, dla wszystkich liczba poziomych bloków wynosi 0, jedna flaga ma główny kolor pomarańczowy, reszta czerwony.

Ponad połowa (56%) państw z tej grupy, leży w strefie NE. Mimo tak dobrego wyniku, państwa są rozrzucone pośród kontynentów.

W przypadku języka, widzimy 3 mniejsze grupy, dlatego nie możemy stwierdzić, że istnieje, w tej grupie, korelacja pomiędzy wyglądem flagi i języka.

35% państw z tej grupy mają jako główną religię odłamy chrześcijaństwa.

Grupa nr 2:

Przykładowe flagi z tej grupy:

Składa się z 13 flag, dla wszystkich liczba poziomych bloków wynosi 0, a główny kolor to biały.

54% państw z tej grupy znajduje się w Europie.

W tej grupie, można też zauważyć, że głównymi religiami jest chrześcijaństwo i odłamy chrześcijaństwa. Może być to powiązane z tym, że głównym kontynentem jest Europa.

Grupa nr 3:

Wszystkie flagi państw należące do tej grupy:

Składa się z 4 flag, dla wszystkich liczba poziomych bloków wynosi 1, dla trzech flag główny kolor to biały, dla jednej czerwony.

Wszystkie państwa z tej grupy:

- znajdują się w strefie NW,
- ich głównymi religiami są odłamy chrześcijaństwa,
- ich populacja nie przekracza 500 tys. mieszkańców.

Takie wyniki mogą być skutkiem małej populacji w badanej grupie (4 państwa).

Grupa nr 4:

Wszystkie flagi państw należących do tej grupy:

Składa się z 5 flag, dla wszystkich główny kolor to złoty, trzy flagi mają liczbę poziomych bloków równą 3, zaś dwie równą 2.

W tej grupie żadne atrybuty (poza stripes i mainhue) nie wyróżniają się pod względem wartości.

Grupa nr 5:

Przykładowe flagi należące do tej grupy:

Składa się z 11 flag, dla wszystkich liczba poziomych bloków wynosi 3, główny kolor dla większości flag to zielony, dla trzech flag to pomarańczowy.

W tej grupie, możemy zauważyć korelację z położeniem geograficznym.

- 73% państw leży w strefie NE
- 64% państw leży w Afryce

Grupa nr 6:

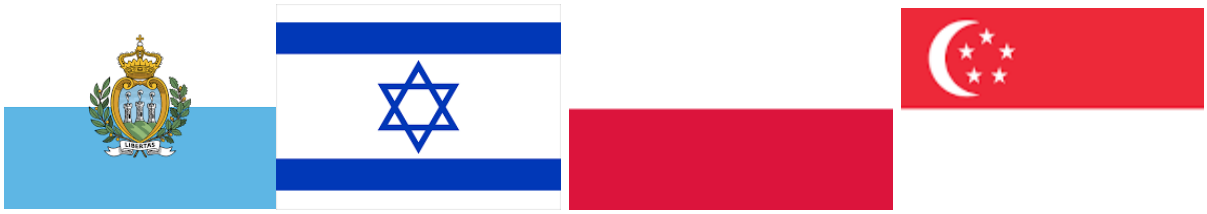
Przykładowe flagi należące do tej grupy:

Składa się z 6 flag, dla wszystkich główny kolor to czerwony, dla większości liczba poziomych bloków wynosi 5, dla jednej flagi 4.

W tej grupie żadne atrybuty (poza stripes i mainhue) nie wyróżniają się pod względem wartości.

Grupa nr 7:

Wszystkie flagi należące do tej grupy:



Składa się z 4 flag, dla wszystkich liczba poziomych bloków wynosi 2, a główny kolor to wszystkich to biały.

Wszystkie państwa z tej grupy, leżą w tej samej strefie – NE.

Co ciekawe, każda z tych flag, posiada dokładnie 2 kolory.

Takie wyniki mogą być skutkiem małej populacji w badanej grupie (4 państwa).

Grupa nr 8:

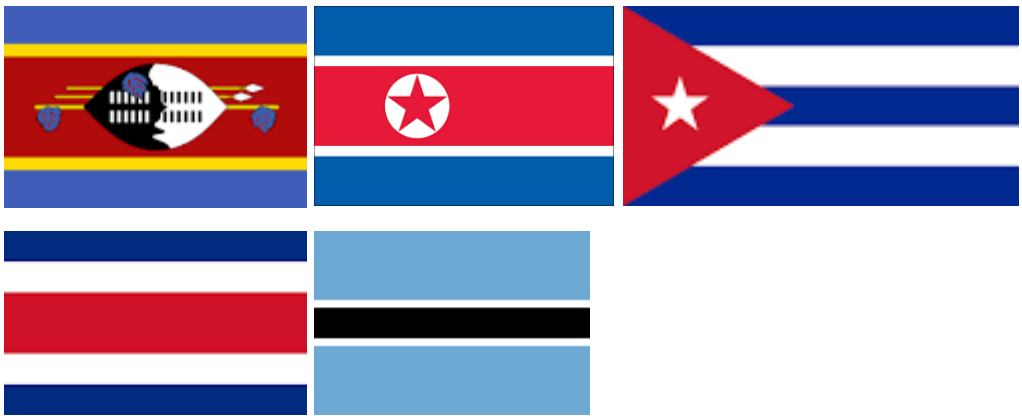
Przykładowe flagi należące do tej grupy:

Składa się z 29 flag, dla wszystkich główny kolor to czerwony, liczba poziomych bloków dla większości wynosi 3, dla dziewięciu flag wynosi 2.

59% państw z tej grupy leży w strefie NE.

Grupa nr 9:

Wszystkie flagi należące do tej grupy:



Składa się z 5 flag, dla wszystkich liczba poziomych bloków wynosi 5, a główny kolor to wszystkich to niebieski.

W tej grupie żadne atrybuty (poza stripes i mainhue) nie wyróżniają się pod względem wartości.

Grupa nr 10:

Wszystkie flagi należące do tej grupy:

Składa się z 3 flag, dla wszystkich liczba poziomych bloków wynosi 3, a główny kolor to wszystkich to czarny.

Wszystkie państwa z tej grupy znajdują się w strefie NE.

Takie wyniki mogą być skutkiem małej populacji w badanej grupie (3 państwa).

Grupa nr 11:

Przykładowe flagi należące do tej grupy:

Składa się z 8 flag, dla wszystkich główny kolor to wszystkich to niebieski, dla większości liczba poziomych bloków wynosi 3, dla jednego wynosi 2.

63% państwa z tej grupy:

- leżą na tym samym kontynencie – Ameryka Północna,
- leżą w tej samej strefie – NW,
- mają ten sam język – hiszpański,
- jako główną religię mają chrześcijaństwo (pozostałe państwa – odłamy chrześcijańskie).

Grupa nr 12:

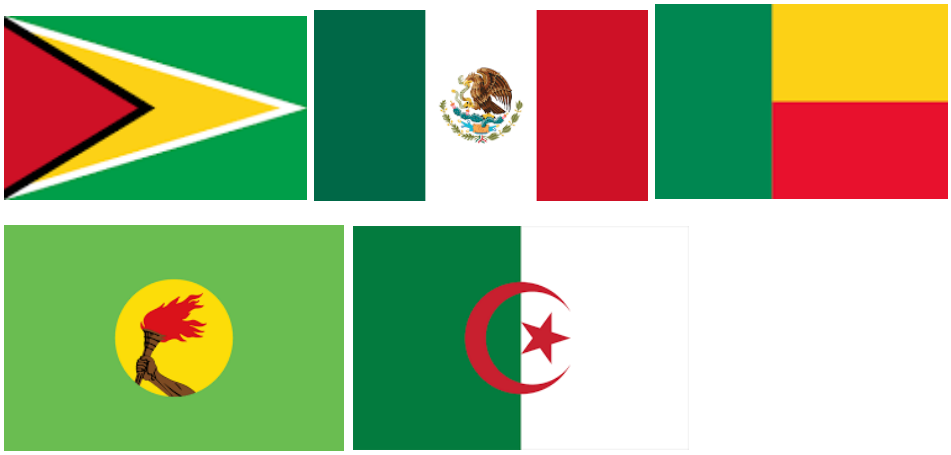
Przykładowe flagi należące do tej grupy:

Składa się z 28 flag, dla wszystkich liczba poziomych bloków wynosi 0, a główny kolor to większości to niebieski, dla dwóch flag czarny.

Mimo tego, że państwa z tej grupy rozrzucone są po całym świecie, to i tak można zauważyć, że głównym (57%) językiem jest język angielski, a główną (68%) religią są odłamy chrześcijaństwa.

Grupa nr 13:

Przykładowe flagi należące do tej grupy:



Składa się z 21 flag, dla wszystkich liczba poziomych bloków wynosi 0, a główny kolor to wszystkich to zielony.

W tej grupie można zauważyć, że znacząca część (48%) państw leży w Afryce.

Widać również, że 43% państw z tej grupy, jest krajami muzułmańskimi.

Grupa nr 14:

Przykładowe flagi należące do tej grupy:

Składa się z 14 flag, dla wszystkich liczba poziomych bloków wynosi 0, a główny kolor dla większości to złoty(żółty), dla dwóch to brązowy.

71% państw z tej grupy, leży w strefie NE

Grupa nr 15:

Wszystkie flagi należące do tej grupy:

Składa się z 2 flag, główny kolor to dla wszystkich to złoty(żółty), dla jednej liczba pionowych bloków wynosi 5, dla drugiej 6.

Wszystkie państwa z tej grupy:

- leżą na tym samym kontynencie – Afryka
- mają jedną religię – religia miejscowa
- przyjmują tą samą wartość jako język – inny

Takie wyniki najprawdopodobniej są skutkiem małej populacji w badanej grupie (2 państwa).

Grupy nr 16-22:

Te grupy zawierają po jednej fladze, więc zostały przez nas przeanalizowane jako jedna grupa państw.

W tych grupach żadne atrybuty nie wyróżniają się pod względem wartości (najprawdopodobniej z powodu tego, że są to państwa których flagi nie pasowały do żadnej innej grupy).

Grupa nr 16:

Składa się z 1 flagi, liczba poziomych bloków wynosi 5, a główny kolor to zielony.

Grupa nr 17:

Składa się z 1 flagi, liczba poziomych bloków wynosi 7, a główny kolor to zielony.

Grupa nr 18:

Składa się z 1 flagi, liczba poziomych bloków wynosi 9, a główny kolor to niebieski.

Grupa nr 19:

Składa się z 1 flagi, liczba poziomych bloków wynosi 14, a główny kolor to czerwony.

Grupa nr 20:

Składa się z 1 flagi, liczba poziomych bloków wynosi 13, a główny kolor to biały.

Grupa nr 21:

Składa się z 1 flagi, liczba poziomych bloków wynosi 11, a główny kolor to czerwony.

Grupa nr 22:

Składa się z 1 flagi, liczba poziomych bloków wynosi 9, a główny kolor to biały.

4 Podsumowanie projektu

Uzyskane przez nas efekty grupowania nas satysfakcjonują. Wydaje nam się, że nasze grupy są wizualnie do siebie podobne. Jesteśmy jednak świadomi, że grupy można wciąż udoskonalać. Gdybyśmy poddali kolejnej analizie poszczególne grupy (szczególnie te z większą ilością flag jak i te, dla których ilość poziomych bloków wynosiła 0) moglibyśmy przy pomocy innych atrybutów jeszcze dokładniej wyodrębnić pewne cechy. Niemniej jednak jesteśmy zdania, że wybrane przez nas atrybutu i sposób grupowania w dobry sposób przedstawiają różnice i podobieństwa między flagami.

Sama praca nad projektem pozwoliła nam zmierzyć się z ‘rzeczywistym’ zbiorem danych, który nie zawsze dawał nam podręcznikowe wyniki. Dzięki temu mogliśmy jeszcze bardziej zagłębić się w temat analizy danych i poszerzyć w dodatkowy sposób swoją wiedzę w tym zakresie.