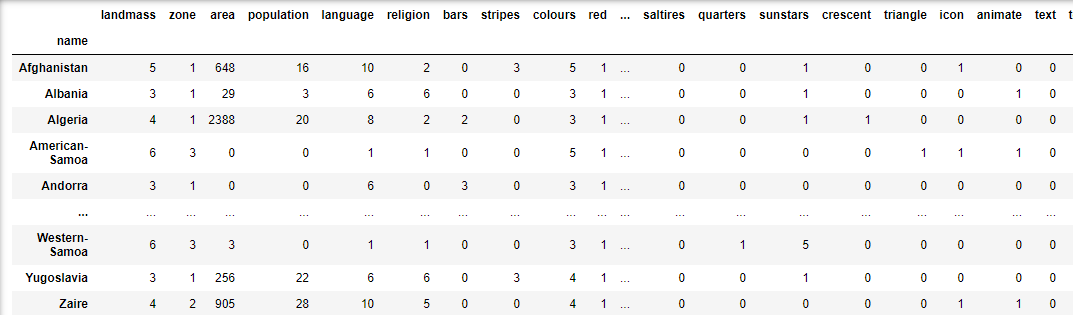
**Podstawy analizy i reprezentacji danych projekt “Grupowanie państw na podstawie wyglądu ich flag”**

Artur Prasuła, Aleksandra Kowalczyk, Kacper Achramowicz

DISCLAIMER: EDYTUJEMY TEN PLIK CZCIONKA CONSTANTIA 14 DO PISANIA (bo lubię ją xD)

1. **Analiza eksploracyjna 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.



Zrzut 1 Ramka danych zawierający zbiór z pliku

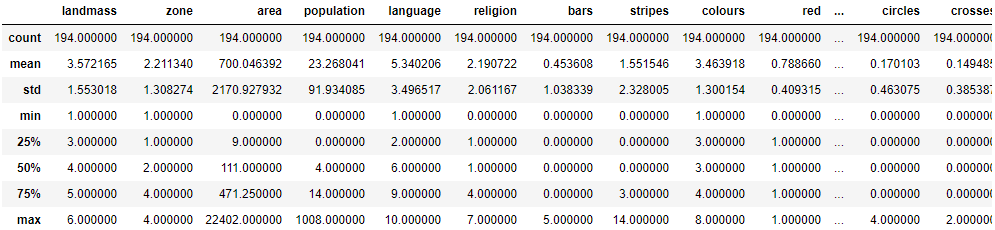
Następnym krokiem było sprawdzenie, czy dane posiadają jakieś braki. Wnioskiem z tego testu było, że zbiór jest pełny. Wynik pokrył się z umieszczonym na stronie <https://archive.ics.uci.edu/ml/datasets/Flags>.

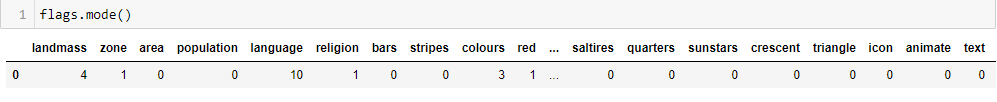
Obraz zawierający stół

Opis wygenerowany automatycznie

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

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).





Następnie utworzyliśmy wykresy punktowe dla każdej pary atrybutów i histogramu częstotliwości występowania dla każdego atrybutu.

Obraz zawierający shoji, budynek, sąsiadująco

Opis wygenerowany automatycznie

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.

Analiza histogramów również pozwoliła nas 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
* 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

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

Obraz zawierający tekst

Opis wygenerowany automatycznie

Możemy zauważyć silne powiązana szczególnie między atrybutami opisującymi kontynent, strefę, powierzchnię, liczebność, religię i język.

* 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 atrubutem ikony, animacje, tekst
* Powyższy 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 mapie, 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 mapie, 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 mapie, 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 aniamcjami – 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

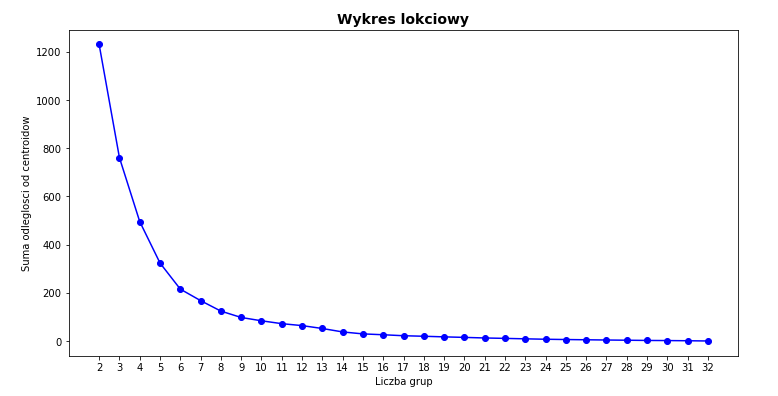
1. **Grupowanie danych**

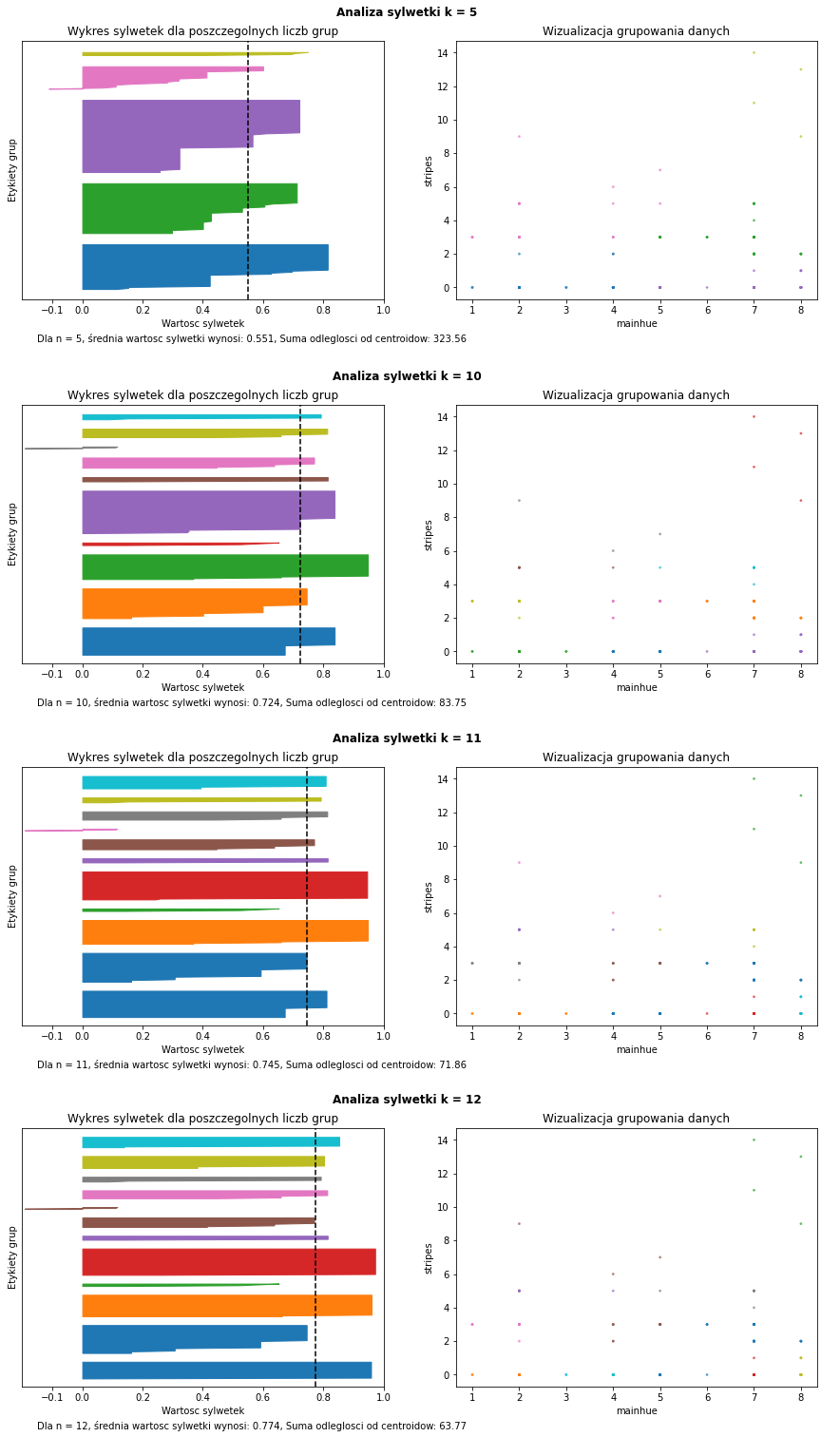
O wybranych atrybutach, wybierzemy metode która spelnia nasze oczekiwania

Dane grupowaliśmy przy użyciu dwóch metod:

* Metody k-średnich
* Metody hierarchicznego grupowania aglomeracyjnego
  1. **Metoda k-średnich**

Pierwszą metodą, którą użyliśmy, była metoda k-średnich. Metoda ta charakteryzuje się tym, że ilość grup ustawiana jest odgórnie. Dlatego wykonaliśmy analizę dla liczby grup z przedziału 2-32. Odpowiednią wynikową liczbę grup wybraliśmy analizując wykres sylwetek dla poszczególnych grup oraz wykres łokciowy.

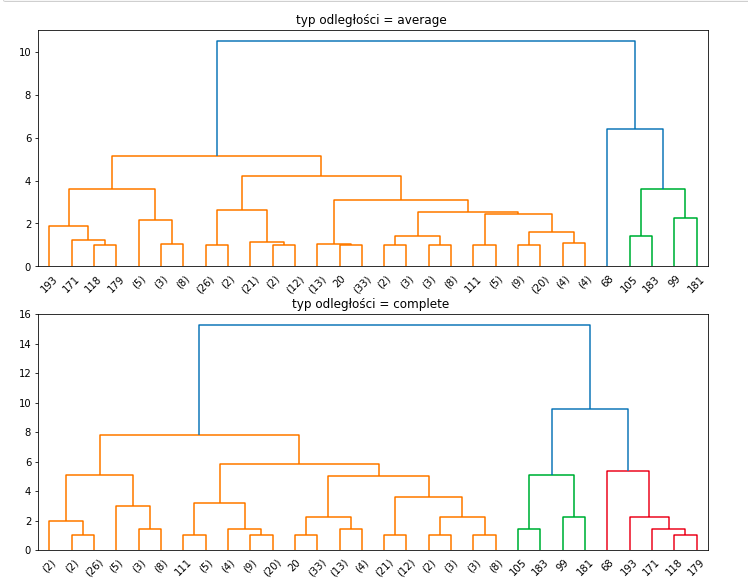


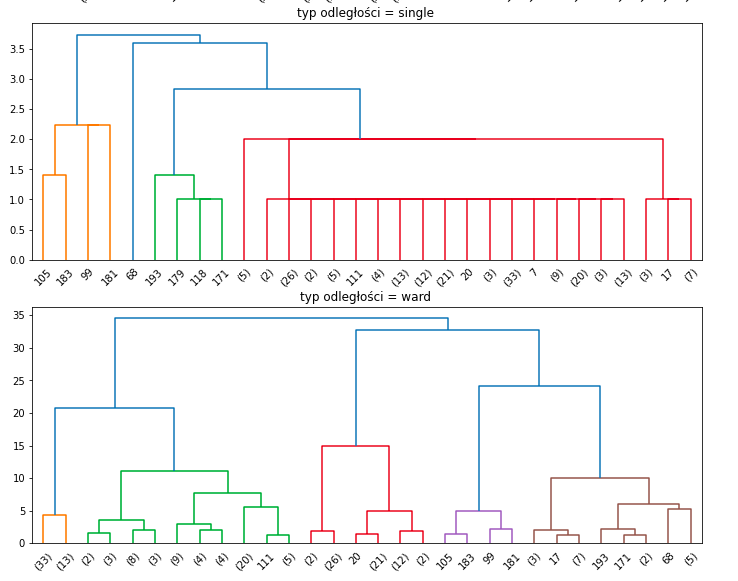


**Wnioski o ilości grup:**

* 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.
  1. **Hierarchiczne grupowanie aglomeracyjne**

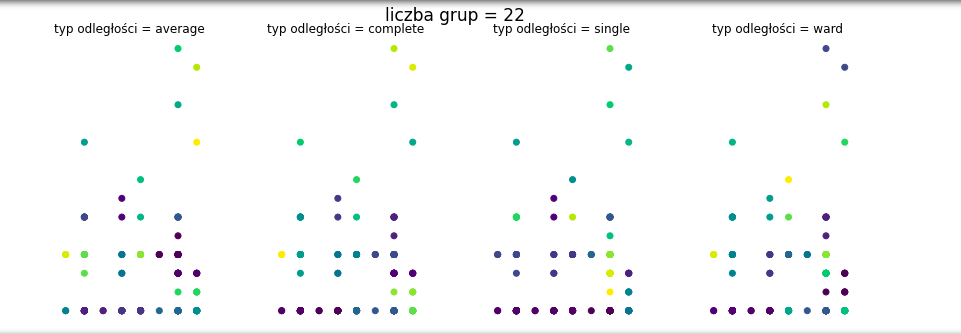
Pracę nad hierarchicznym grupowaniem aglomeracyjnym rozpoczęliśmy od analizy i wyboru sposobu określenia odległości między skupiskami obiektów. W tym celu stworzyliśmy 4 dendrogramy, każdy dla jednego typu odległości.





Na początku wykluczyliśmy metodę 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 metodę average, ponieważ wiele różnych obiektów w tej metodzie zostało potraktowanych jako podobne i połączone w jedną grupę, przez co otrzymaliśmy bardzo uśrednione wyniki. Metoda ward i complete dawała bardzo podobny wynik – podobną liczbę grup i wygląd dendrogramu. Analizując jednak grupy doszliśmy do wniosku, że lepsze przyporządkowanie daje nam metoda 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.



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

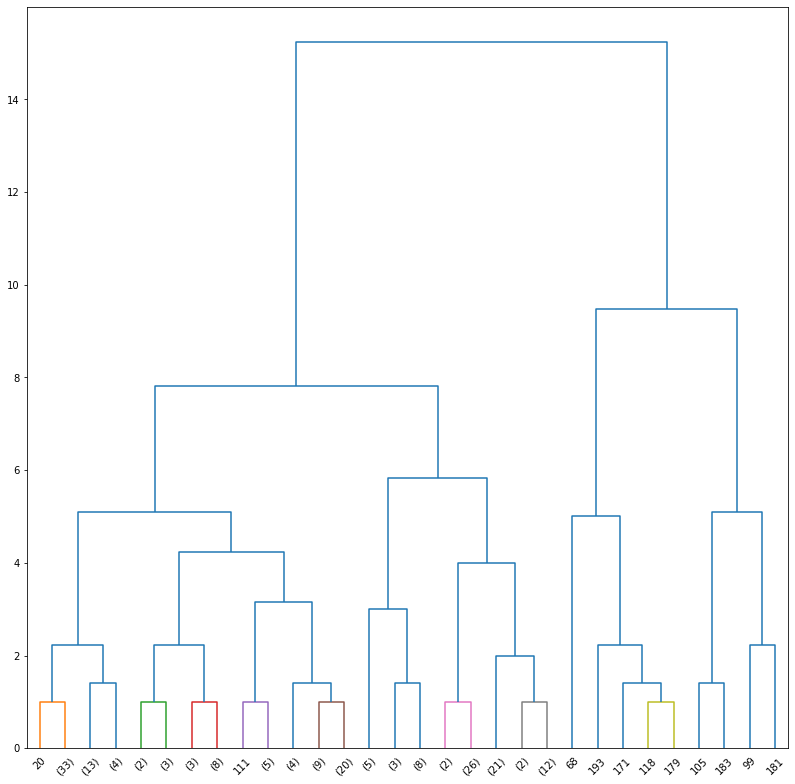
Wykresy potwierdziły naszą decyzje o wykluczeniu metody single- widzimy, że tworzy przede wszystkim dwie grupy o bardzo dużej ilości różnych obiektów.

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 sposobu ‘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.

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.



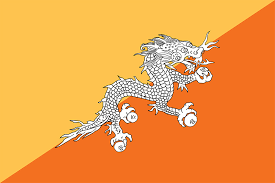
* 1. **Porównanie obu metod grupowania**

Po krótkiej analizie wyników doszliśmy do wniosku, że pierwszy sposób grupowania jest o wiele mniej efektywny, ponieważ niektóre flagi z jednej grupy są do siebie bardzo 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:

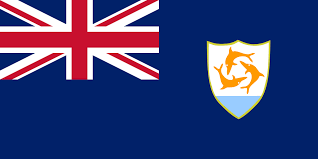
  

*Flaga 1: Angola Flaga 2: Boliwia Flaga 3: Chile*

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: Anguila Flaga 8: Cypr Flaga 9: Irlandia

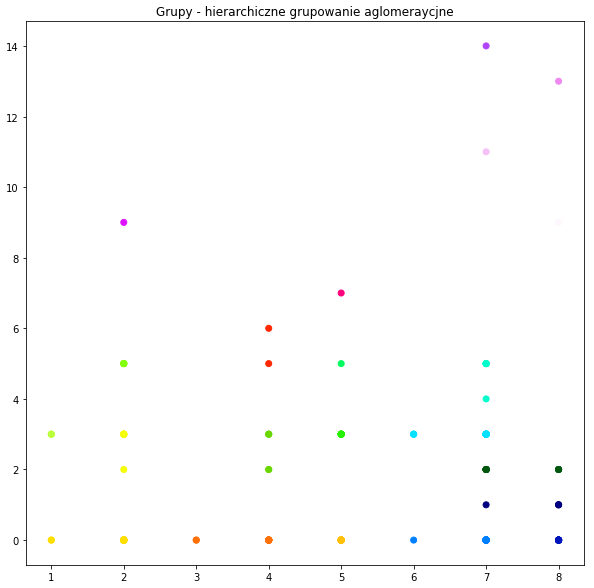
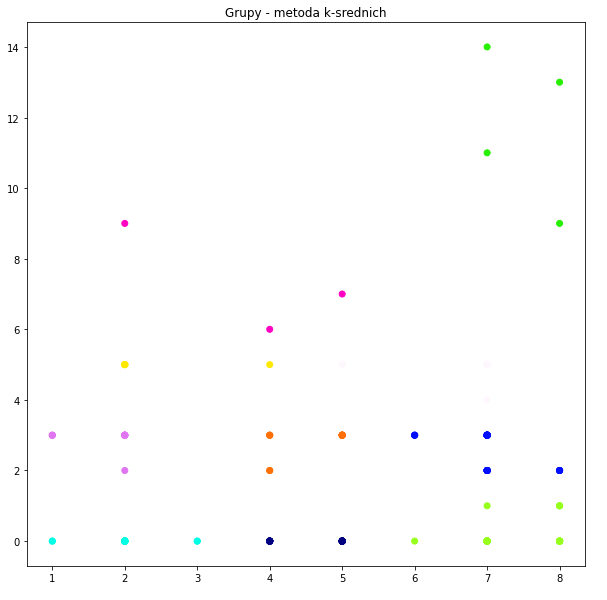
Przykładowe flagi z grupy nr 10 w hierarchicznym grupowaniu:

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, natomiast już grupa numer 10 zawiera flagi znacznie się różniące, gdzie druga metoda dalej grupuje w sposób akceptowalny.

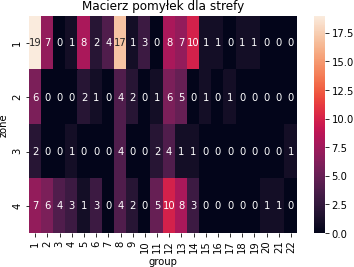
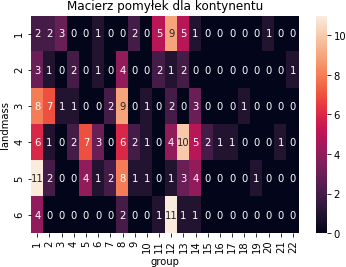
Wykresy grup dla obu metod pokazują, że metoda k-średnich umieszcza częściej w grupach flagi mające różne ilości poziomych pasków jak i mające inny kolor przewodni. Na tej podstawie można stwierdzić, że to kolejny argument pokazujący wyższość metody grupowania aglomeracyjnego w rozpatrywanym przez nas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

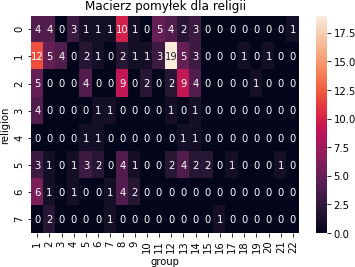
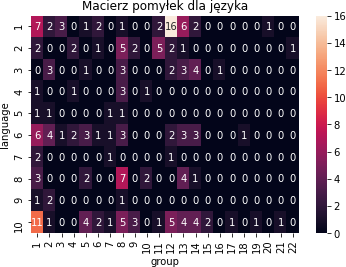


1. **Wnioski z pogrupowania**

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.

**Wizualizacja i analiza efektów grupowania:**

**Grupa nr 1:**

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:**

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 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:**

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

**Grupa nr 5:**

W tej grupie, możemy zauważyć korelację z położeniem geograficznym.

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

**Grupa nr 6:**

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

**Grupa nr 7:**

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:**

59% państw z tej grupy leży w strefie NE.

**Grupa nr 9:**

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

**Grupa nr 10:**

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:**

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:**

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:**

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:**

71% państw z tej grupy, leży w strefie NE

**Grupa nr 15:**

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).

#TO DO:

5 FLAG DLA KAŻDEJ GRUPY(DO GRUPY NR 16)