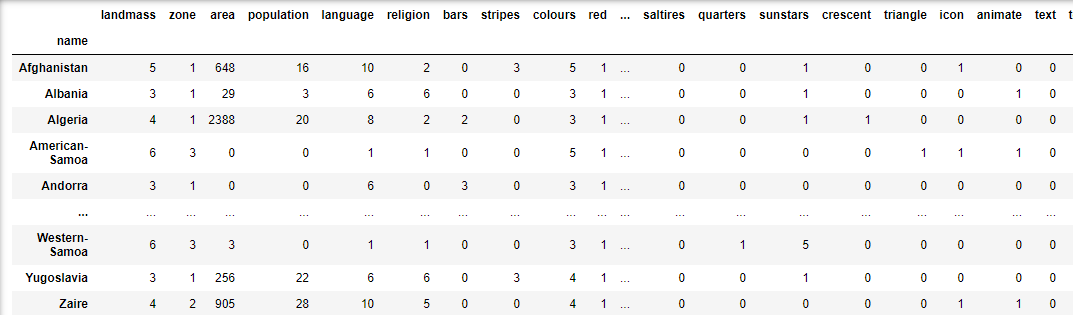
**Podstawy analizy i reprezentacji danych projekt “Grupowanie państw na podstawie wyglądu ich flag”**

Artur Prasuła, Aleksandra Kowalczyk, Kacper Achramowicz

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

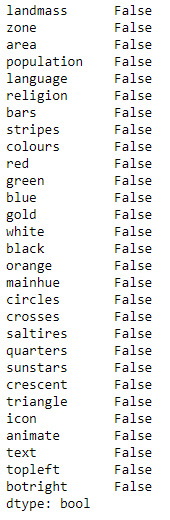
1. **Analiza eksploracyjna zbioru danych**

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 kolory na liczby dziesiętne.



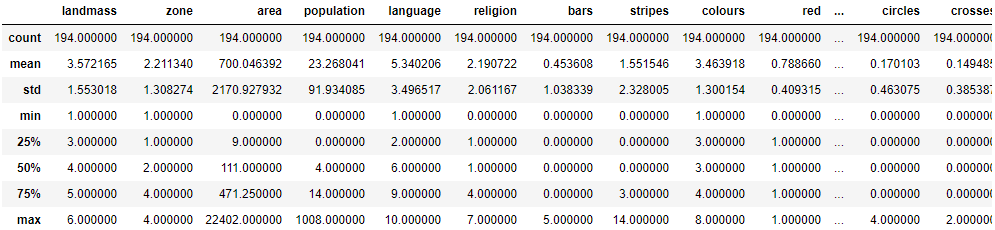
Zrzut 1 Ramka danych zawierający zbiór z pliku

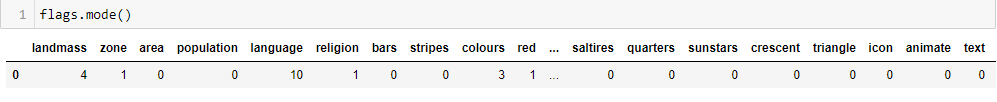
Następnym krokiem było sprawdzenie, czy dane posiadają jakieś braki. Wnioskiem z prostego testu było, że zbiór jest pełny. Wynik pokrył się z umieszczonym na stronie <https://archive.ics.uci.edu/ml/datasets/Flags>.



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 pionowych słupków we fladze).



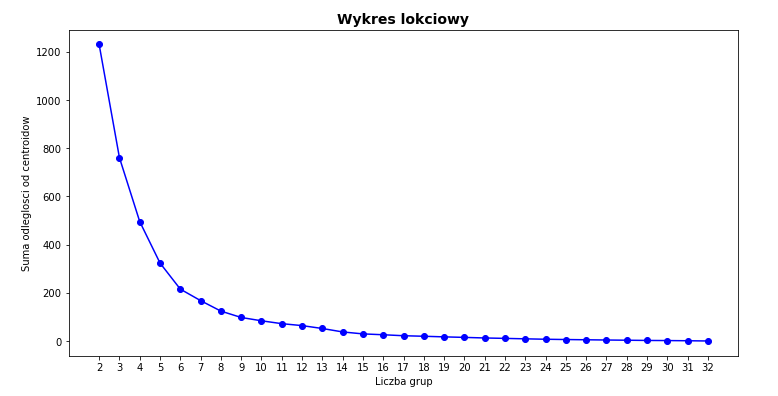


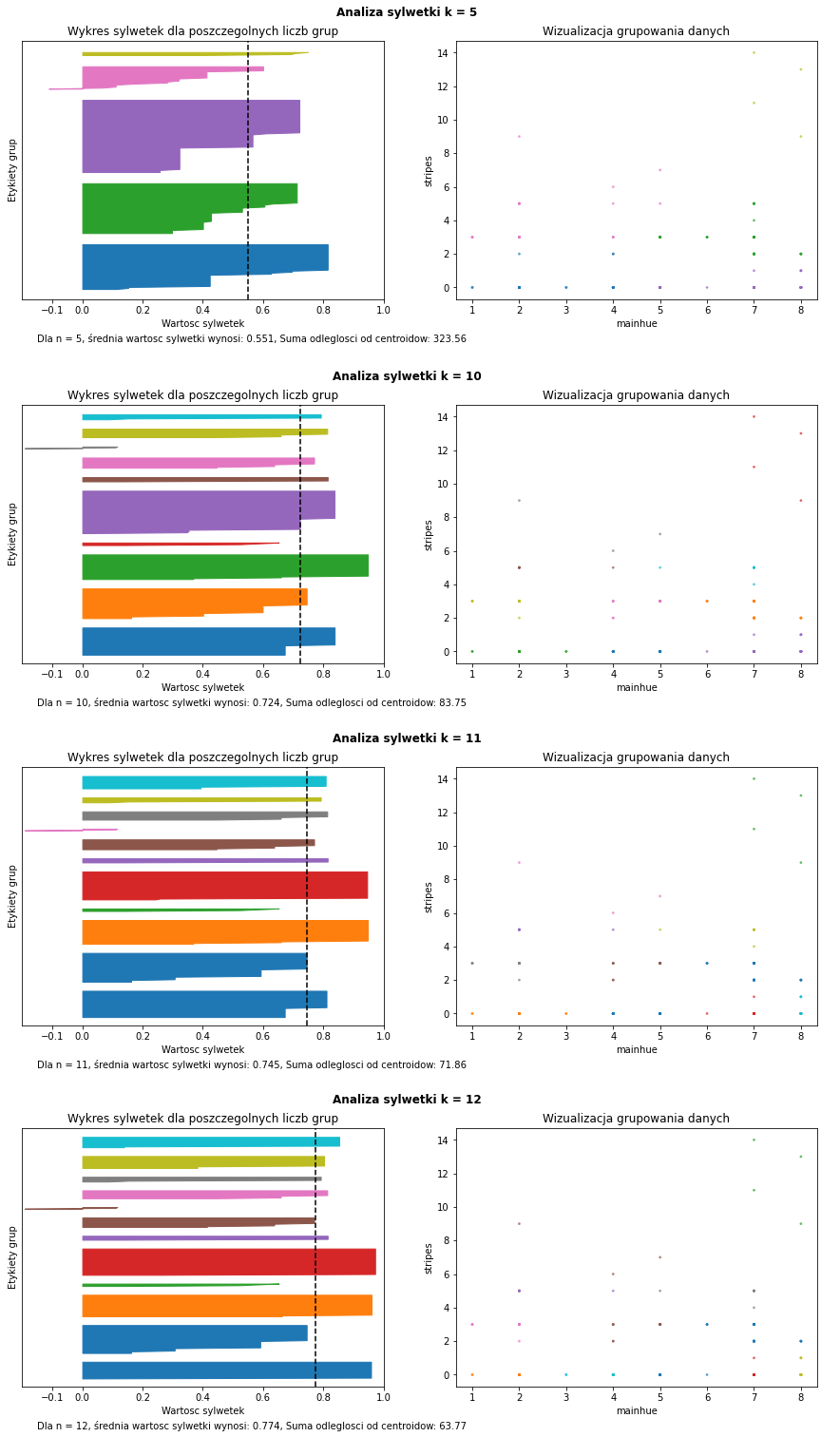
1. **Grupowanie danych**

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 2-22. Odpowiednią wynikową liczbę grup wybraliśmy analizując wykres sylwetek dla poszczególnych grup oraz wykres łokciowy.





**Wnioski o ilości grup:**

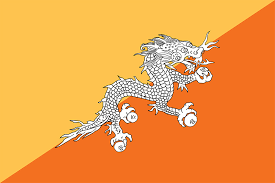
* Wykres łokciowy łagodnie zwalnia, dlatego ciężko znaleźć optymalną liczbę grup.
* Dla ilości grup 6-9 widać, że „spadek” wykresu maleje.
* Z wykresu łokciowego wynika, że optymalna ilość grup mieści się w przedziale 10-12 (włącznie).
* Z analizy sylwetki wynika, że optymalna ilość grup jest równa 5.
* 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**
  2. **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 na utworzenie aż 32 grup, których ilość potem zmniejszyliśmy do 22 ze względu na dużą ilość grup zawierających tylko jedną flagę. W tym wypadku większa ilość grup drugiej metody pozwala na większą dokładność.

Przykładowe flagi z grupy nr 1 w pierwszej metodzie:

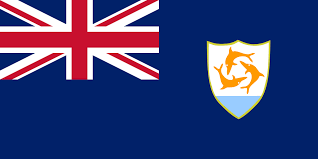
  

*Flaga 1: Angola Flaga 2: Boliwia Flaga 3: Chile*

Przykładowe flagi z grupy nr 1 w drugiej metodzie:  

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

Przykładowe flagi z grupy nr 10 w pierwszej metodzie:

Flaga 7: Anguila Flaga 8: Cypr Flaga 9: Irlandia

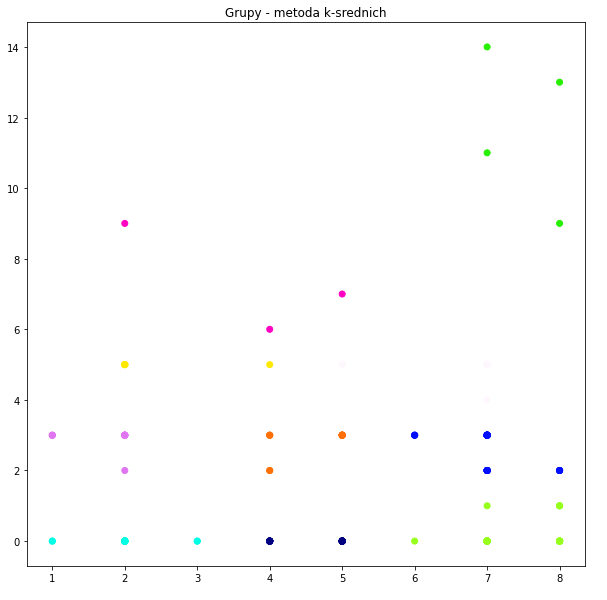
Przykładowe flagi z grupy nr 10 w drugiej metodzie:

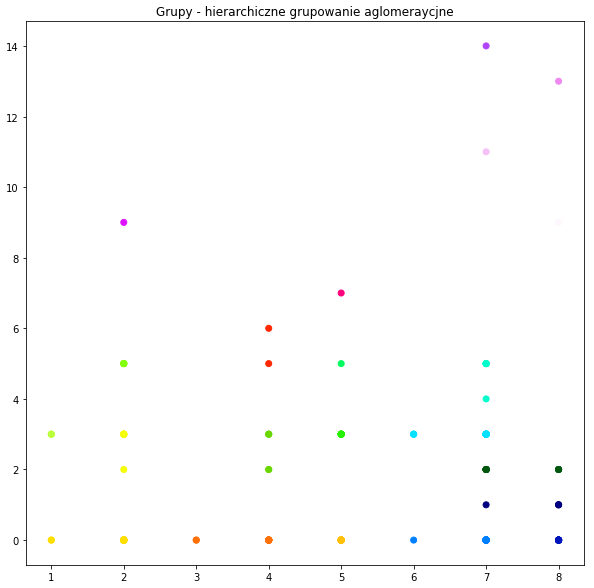
  

Flaga 10: RFN Flaga 11: Jordania Flaga 12: Egipt

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 w grupach flagi mające różne ilości poziomych pasków jak i mające inne kolor przewodni. Na tej podstawie można stwierdzić, że to kolejny argument pokazujący wyższość metody grupowania aglomeracyjnego w rozpatrywanym przez nas problemie.

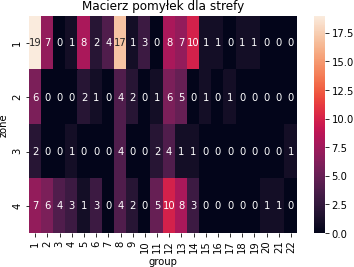
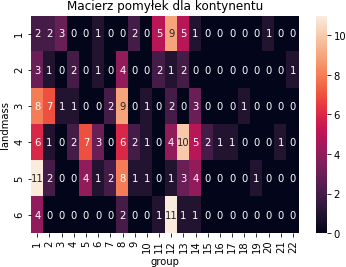


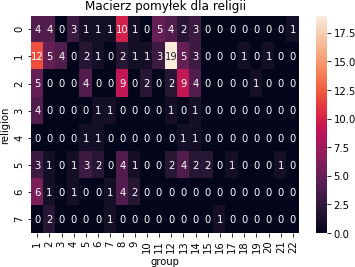
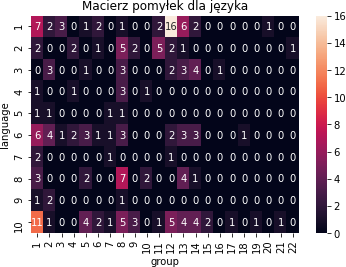


1. **Wnioski z pogrupowania**

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 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 – Ethnic //TO DO
* 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 jednym państwie, więc zostały przez nas przeanalizowane jako jedna grupa państw, które nigdzie nie pasowały.

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)