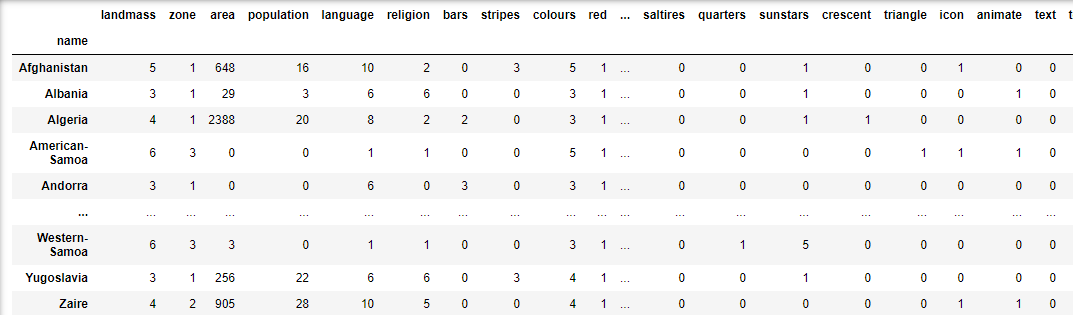
**Podstawy analizy i reprezentacji danych projekt “Grupowanie państw na podstawie wyglądu ich flag”**

Artur Prasuła, Aleksandra Kowalczyk, Kacper Achramowicz

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

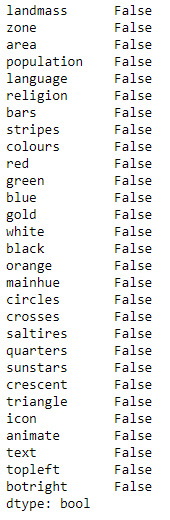
1. **Analiza eksploracyjna zbioru danych**

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.



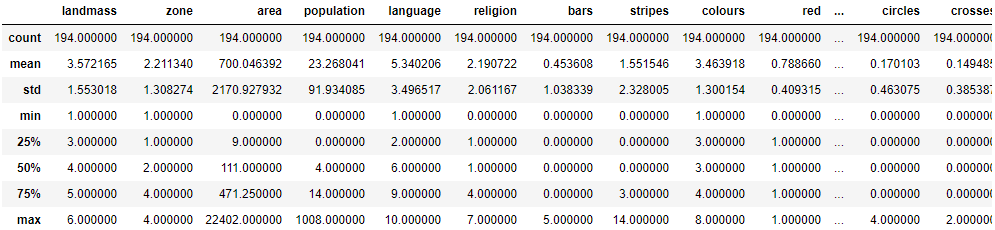
Zrzut 1 Ramka danych zawierający zbiór z pliku

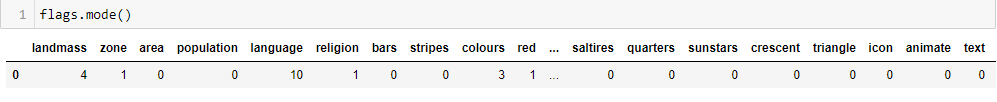
Następnym krokiem było sprawdzenie, czy dane posiadają jakieś braki. Wnioskiem z prostego testu było, że zbiór jest pełny. Wynik pokrył się z umieszczonym na stronie <https://archive.ics.uci.edu/ml/datasets/Flags>.



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:

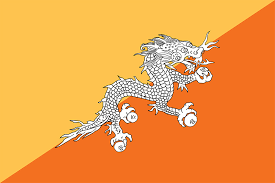
* Metody k-średnich
* Metody hierarchicznego grupowania aglomeracyjnego

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:

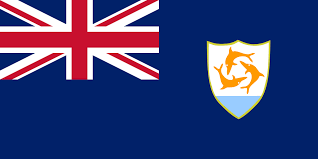
  

*Flaga 1: Angola Flaga 2: Boliwia Flaga 3: Chile*

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.