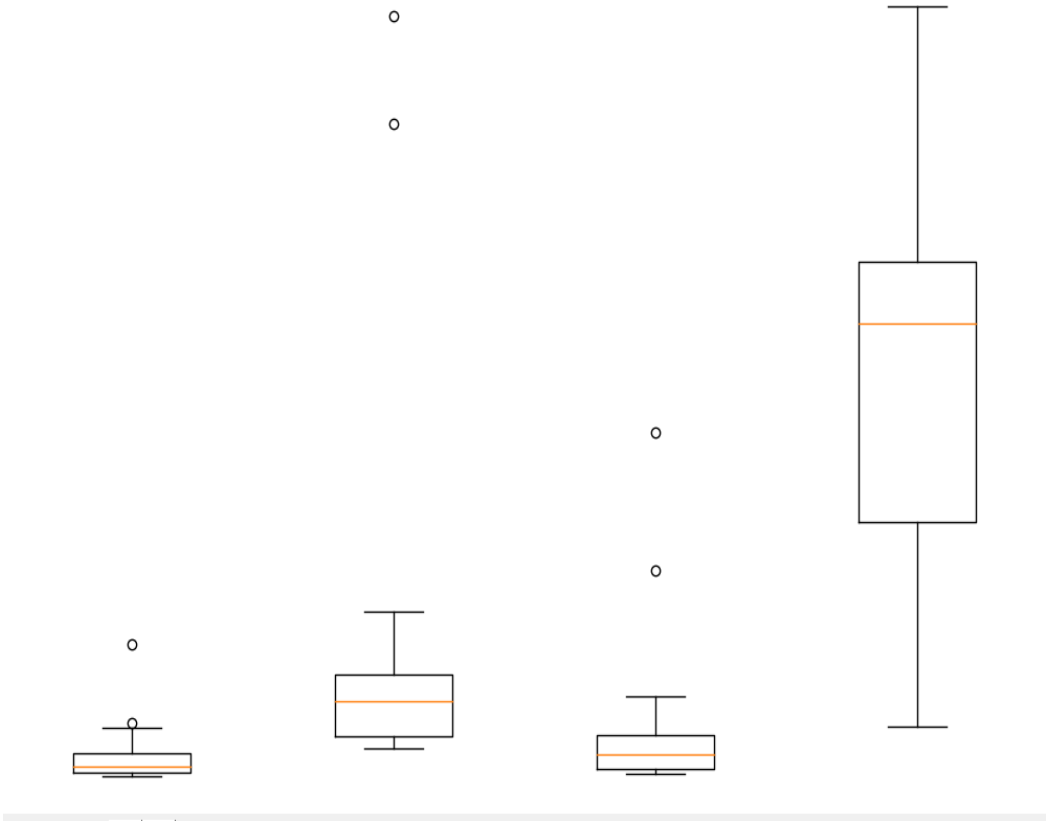
Klasterovanje

- **Članovi tima:**
  Panić Miloš SW19/2018
  Bjelica Luka SW21/2018
  Šerbedžija Luka SW32/2018

- **Pristup problemu:**

Prvi korak u rešavanju problema je bio analiza datog skupa podataka. U pitanju je skup sa 4 obeležja od kojih su 2 kategorička. Kategorije su relativno balansirane, s obzirom da se podaci zasnivaju na postojećim državama 1970tih pa je broj veoma ograničen. Na sledećoj slici su prikazani boxplotovi zarade datih regiona i to redom: Africa, Americas, Asia, Europe



Sa slike se može zaključiti da Europe ima znatno veći prosek plata u odnosu na ostale kontinente koje imaju slične proseke. Takođe se mogu uočiti i outlieri. Većina outliera ima značajan uticaj na model tako da nisu izbačeni svi outlieri.

- **Rad sa treningu skupom podataka**

Pre svega, sva kategorička obeležja je potrebno labelirati. U tu svrhu smo koristili label encoding i to za oil: 0 - ne, 1 - da i 0, 1, 2, 3 za redom Africa, Americas, Asia, Europe.

Trening skup podataka sadrži neke nedostajuće vrednosti. Pokušali smo ih popuniti koristeći više tehnika:

1. Kompletno zanemarivanje redova koji imaju nedostajući podatak
2. Korišćenjem ostalih redova kojima ne fale vrednosti uz pomoć srednje vrednosti i medijane
3. Korišćenjem vrednosti koje smo pronašli na internetu za prosečan broj umrlih na 1000 novorođenih za date regione

Nad ulaznim podacima je takođe pokušana i normalizacija. Evaluirali smo rešenja gde bi samo jedna kolona bila normalizovana, kao i rešenje sa obe kolone normalizovane.

Pored normalizacije, pokušane su i neke druge transformacije podataka. Nad kolonom income su primenjene log10, log2, log, sin, cos. Nad kolonom infant je primenjeno celobrojno deljenje i dodavanje jedinice. Ovime smo sve vrednosti podelili u 13 grupa. Ideju smo dobili sa prezentacije gde su podaci vizualizovani uz pomoć histograma.

S obzirom na to da nisu sva obeležja od istog značaja za rešavanje problema, izbacili smo obeležje oil.

- **Treniranje modela:**

Treniranje *Gaussian Mixture* modela se svelo na treniranje raznih kombinacija transformacija trening skupa podataka i vrednosti parametara modela. Parametri modela koje smo pokušali optimizovati su:

n_components, *covariance_type*, *random_state, n_init* i *max_iter*

Zaključak:

Kao najbolje rešenje nam se pokazalo da je u redu zanemariti redove sa nedostajućim vrednostima i da je potrebno primeniti celobrojno deljenje na *infant* kolonu uz jednostavan *Gaussian Mixture* modela sa parametrima:

n_components = 4, random_state =1