DataWarehouse

Meilleur pays pour étudier

SIHAMDI Mostefa, BOUSBA Abdellah

UE BIUM 2020-2021, Encadrantes : Laure Soulier et Agnès Mustar

M1 DAC

Table of Contents

I. Contexte

Plus de 5 millions de personnes font leurs études supérieures en dehors de leur pays d'origine, Cela représente 2 étudiants sur 100. L'augmentation rapide de la mobilité universitaire internationale devrait se maintenir. Dans le monde très interconnecté d'aujourd'hui, la mobilité étudiante est pourtant sensible aux évolutions politiques et économiques mondiales. Par conséquent, choisir une université qui réponde aux attentes de l'étudiant tant sur le plan économique que sur celui de la qualité de l'enseignement devient nécessaire au vu du nombre d'universités dans le monde, dans le cadre du projet BIUM on cherche à répondre à la problématique : « quel est le pays le plus approprié pour continuer ses études supérieures», où nous étudierons, la qualité de l'enseignement, les frais d'inscriptions, les bourses offertes, le coût de la vie(loyer, courses, transport, santé…).

II. Données :

Nous nous sommes appuyés sur plusieurs données pour faire notre analyse que nous allons vous présenter ainsi que le processus d'extraction. Tout d'abord **world.sql**, la base de données qui nous a été fourni et qui contient déjà beaucoup d'information sur un grand ensemble de pays du monde, tels que le nom et leur code.

universities_students.csv : contient le nombre d'étudiants total et étrangers dans chaque université avec leur pays.

universities_scores.csv : contient des université avec leur pays, la qualité d'enseignement, volume, revenus et réputation de la recherche, citation de l'université (influence).

university_fees.csv : contient les frais universitaires par université avec pays, extrait par nous-mêmes en utilisant le web scrapping sur le site https://www.unipage.net/en/

scholarships.csv :contient le nombre de programme de bourse par pays, extrait par nous-mêmes en utilisant le web scrapping sur le site https://scholarship-positions.com/

unemployment_rate.csv : contient le taux de chômage par pays, extrait par nous-mêmes en utilisant le web scrapping sur le site https://en.wikipedia.org/wiki/List_of_countries_by_unemployment_rate

health_cost.csv : contient les frais médicaux par pays.

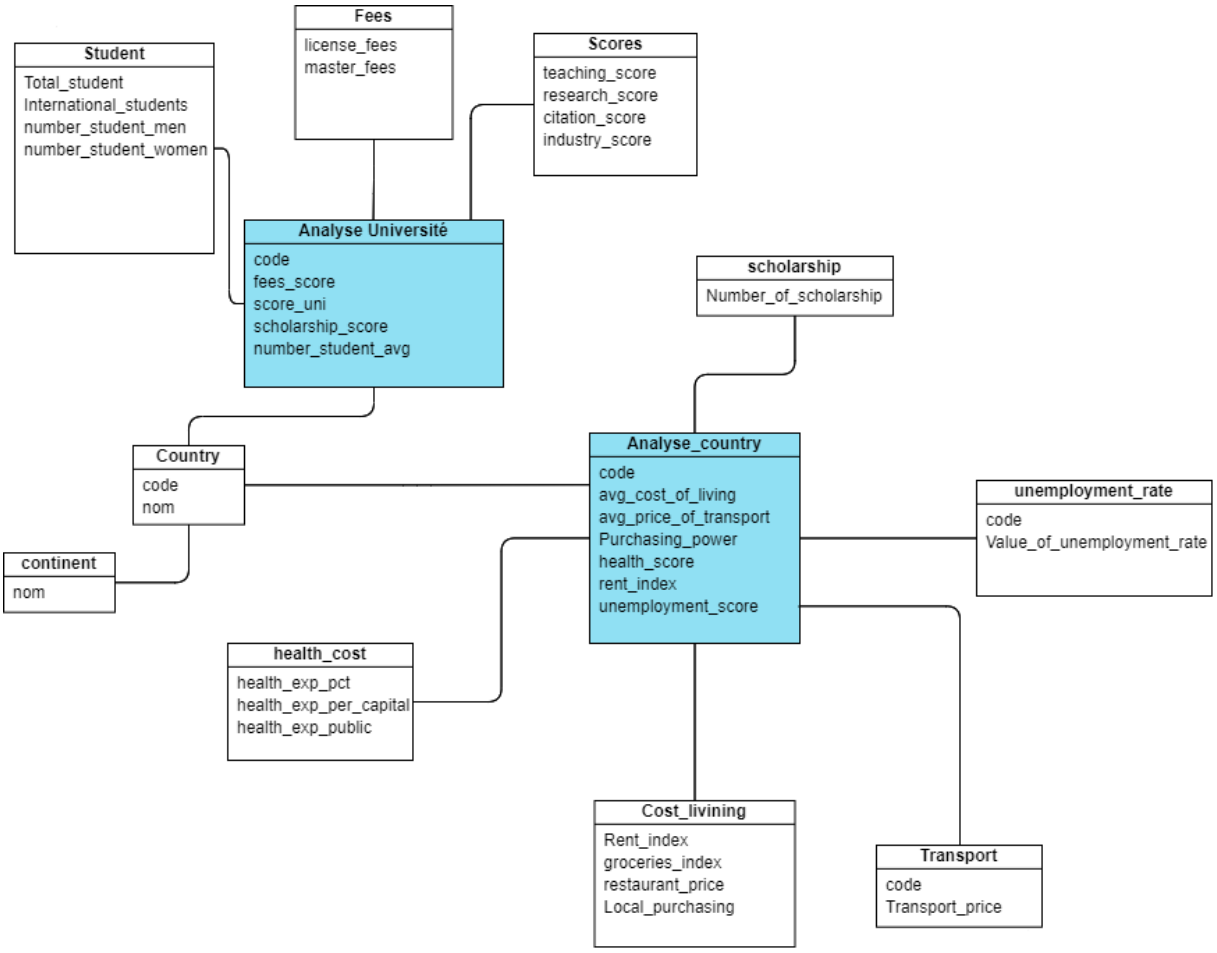
cost_of_living.csv : contient le cout de vie par pays, contient le loyer, les courses, restaurant et pouvoir d'achat.

transport_costs.csv : contient le Prix par pays d'un billet de transport local, extrait par nous-mêmes en utilisant le web scrapping sur le site : https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=18

III. Modélisation :

Dans notre démarche, nous allons diviser notre problématique en deux sous problèmes, le premier est de trouver les métriques associer au cout de vie de chaque pays pour analyser la situation économique et sociale. La deuxième c'est de trouver les métriques associer au niveau d'enseignement de chaque pays en analysant la situation de chaque université. Enfin, nous faisons l'analyse et la combinaison entre les deux résultats pour trouver les pays les plus appropriés pour continuer ses études supérieures.

La figure suivante contient le schéma en constellation de notre datawarehouse :



Après extraction des données dans notre datawarehouse nous obtenons la représentation en logique dénormalisée ci-contre :

Country(code_country,continent,nom)

Health_cost(id_health,health_exp_pct,health_exp_per_capital,health_exp_public)

Transport(id_transpot,prix)

Cost_living(id_cost,rent_index,groceries_index,restaurant_index,Local_purchasing)

Scholarship (id_scholarship,number)

Unemplomyement_rate(id_rate,Taux)

Score(id_score,teaching_score,research_score,citation_score,industry_score)

Fees_univ(id_fees,licence_fees,master_fees)

Students(id_student,international_student,)

Analyse_Universite(id_student, id_score, id_country,id_fees, fees_score, score_uni, scholarship_score, number_student_avg)

Analyse_pays(id_rate,code_country, id_health, id_transpot,id_cost, id_scholarship,avg_cost_of_living, avg_price_of_transport, Purchasing_power, health_score, rent_index, unemployment_score)

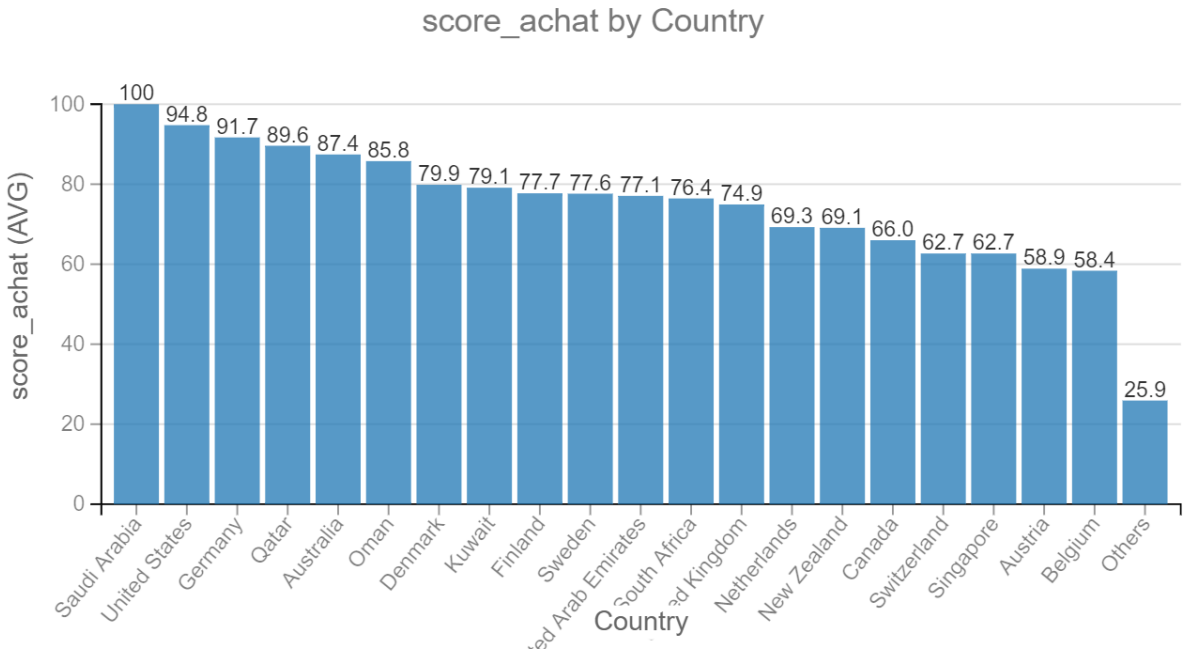
IV. Analyse :

Dans un premier temps, pour répondre à notre problématique, nous avons défini plusieurs métriques qui pourraient permettre de dire quel est le meilleur pays du monde pour y vivre et étudier, les métriques sont calculées à partir des attributs de chaque dimension. Nous avons deux types d'analyse, le premier type est celui qui concerne la situation économique et sociale dans les pays, le second type étudie la qualité de l'enseignement dans les universités pour chaque pays.

1. Analyse pays :

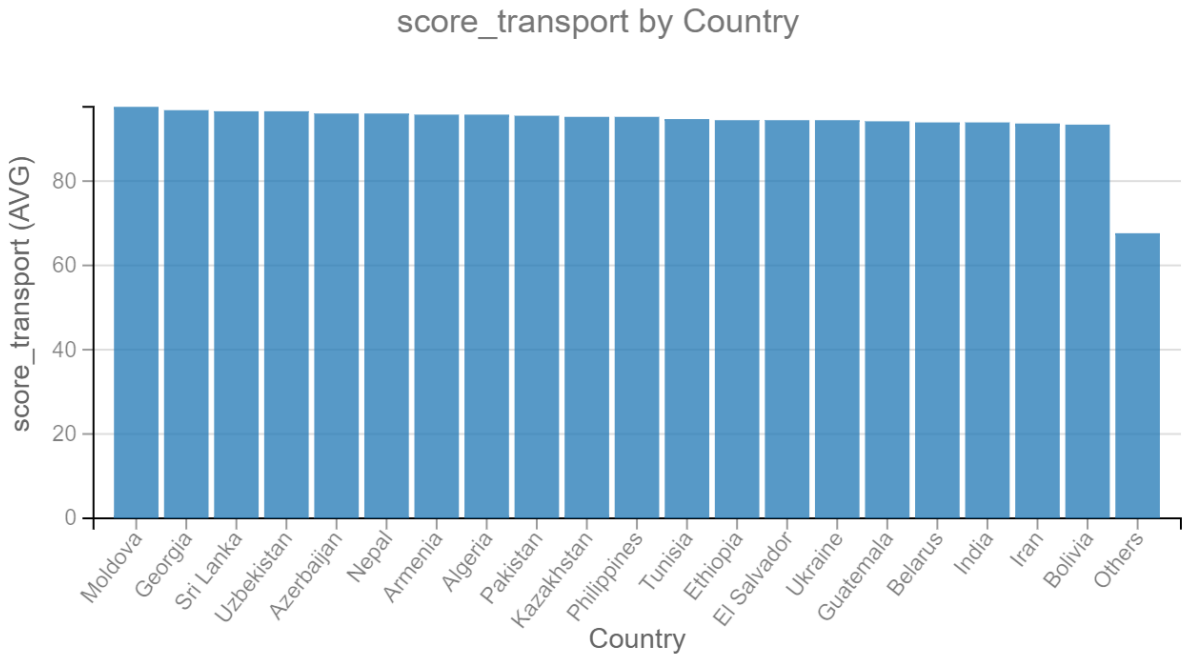
1.1. Pouvoir d'achat :

On commence par le pouvoir d'achat dans chaque pays, pour analyser cette métrique (s'appeler score_achat dans notre modélisation) on a utilisé la dimension cost_livining en faisant une somme pondérée sur les attributs Restaurant_price, local_purchasing, groceries_index, pour trouver un pourcentage par rapport à chaque pays.



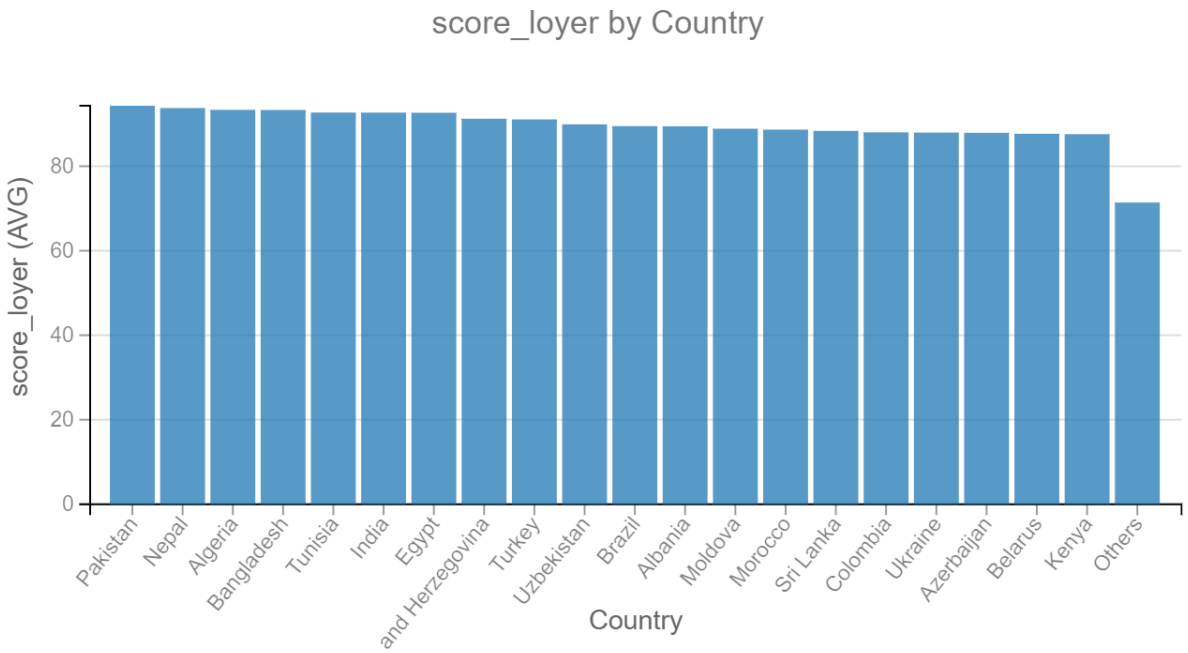
1.2. Coût de transport :

Les moyens de transport sont utilisés par les étudiants quotidiennement pour aller au lieu d'enseignement. En conséquence, nous définissons la mesure prix des transport de chaque pays (score_transport dans notre modélisation).



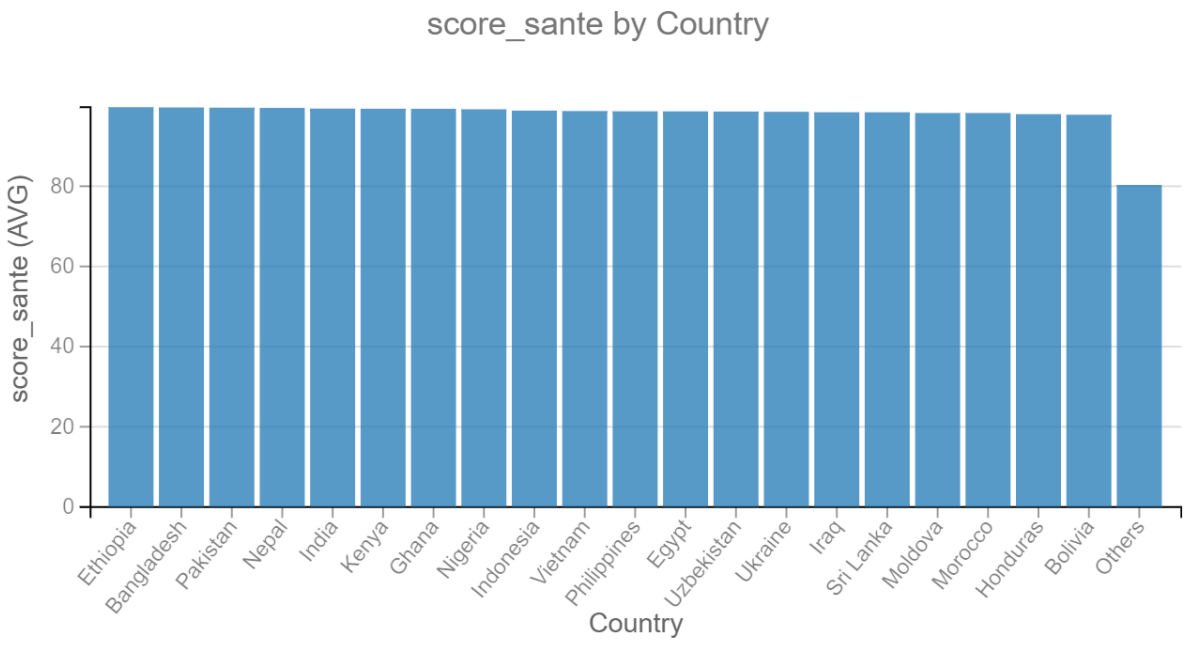
1.3. Coût de loyer :

Le loyer représente une grande partie des dépenses d'argent pour un étudiant. Alors, on analyse le taux des prix de loyer pour chaque pays (en %) (score_loyer dans notre modélisation).



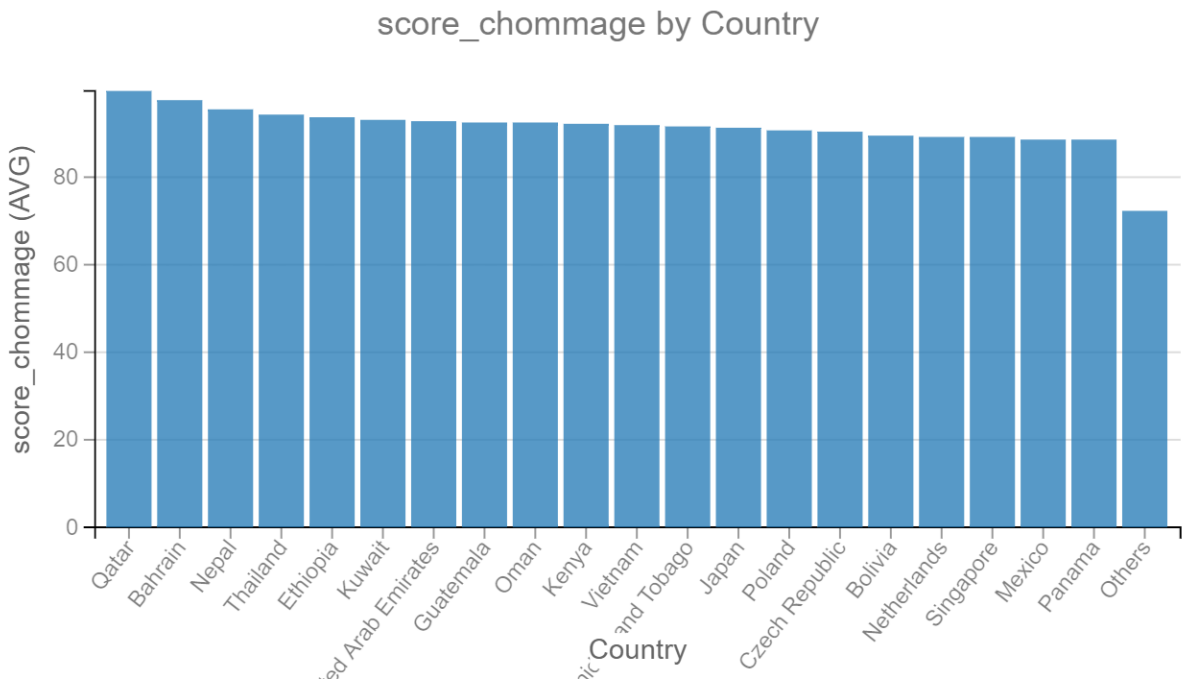
1.4. Frais médicaux :

On fait une étude aussi sur les frais d'une consultation médicale pour chaque pays (en %), on a calculé ce métrique (s'appeler score_sante dans la modélisation) en utilisant la dimension health_cost, et en moyennant les frais de santé public et privé avec les aides proposer par chaque pays.

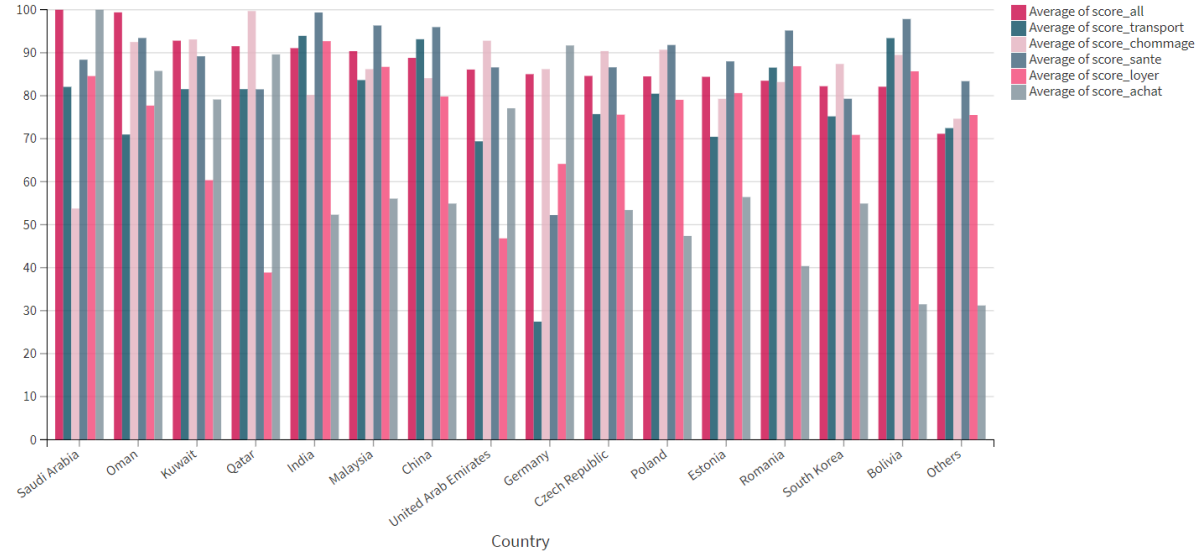


1.5. Taux de chômage :

Comme nous le savons, plusieurs étudiants restent dans leur pays d'étude pour chercher un travail. Donc on va étudier le taux de chômage (en %) (s'appeler score_chomage dans notre modélisation).



Dans l'histogramme suivant on essaye de faire une synthèse sur les analyses précédentes pour chaque pays. Après avoir combiner les scores précédents on a pris les top 20 pays :

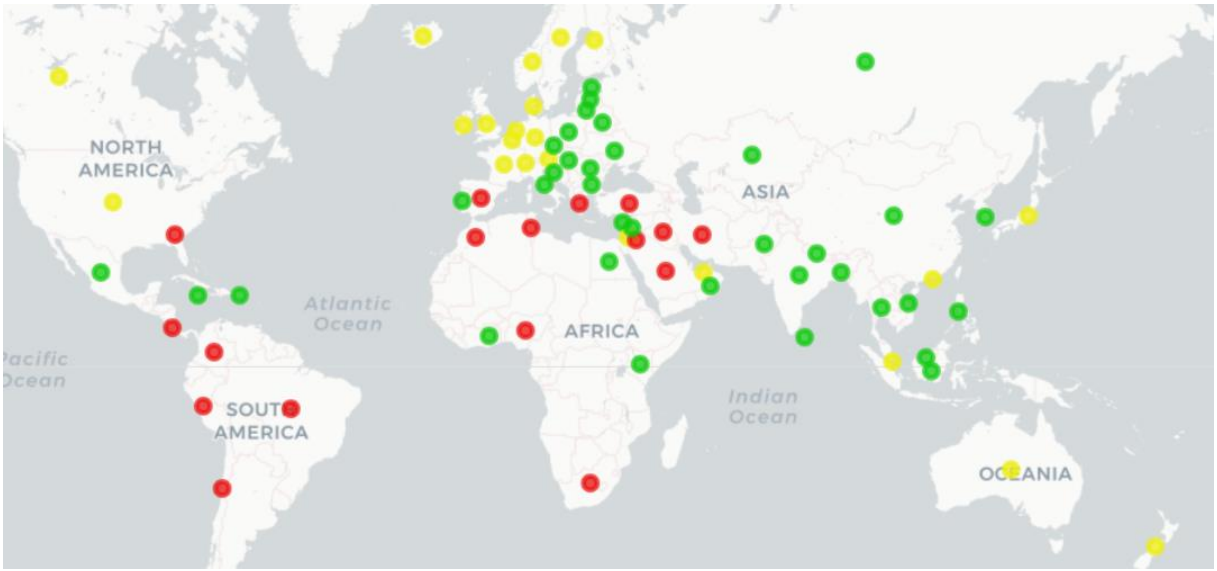


On remarque que la majorité des mesures sont assez équilibrés.

1.6. Clustering :

Nous avons fait un clustering (K-Means) en gardant 3 clusters (et un outlier) sur les cinq métriques score_trasport, score_chomage, score_sante, score_loyer et score_achat.

Grace au clustering nous pouvons identifier 3 types de pays : cluster_0 (mauvais), cluster_1 (bon), cluster_2 (moyen) que nous avons représente sur la carte ci-dessous respectivement en rouge, vert et jaune. Les pays en rouge (cluster_0) ont un pouvoir d'achat très faible et un taux de chômage le plus bas. Les pays en vert (cluster_2) ont un faible pouvoir d'achat mais tous autre mesure est acceptable. Enfin en jaune (cluster_1) on remarque un équilibre entre les mesures et un pouvoir d'achat très élevé et peu de chômage.

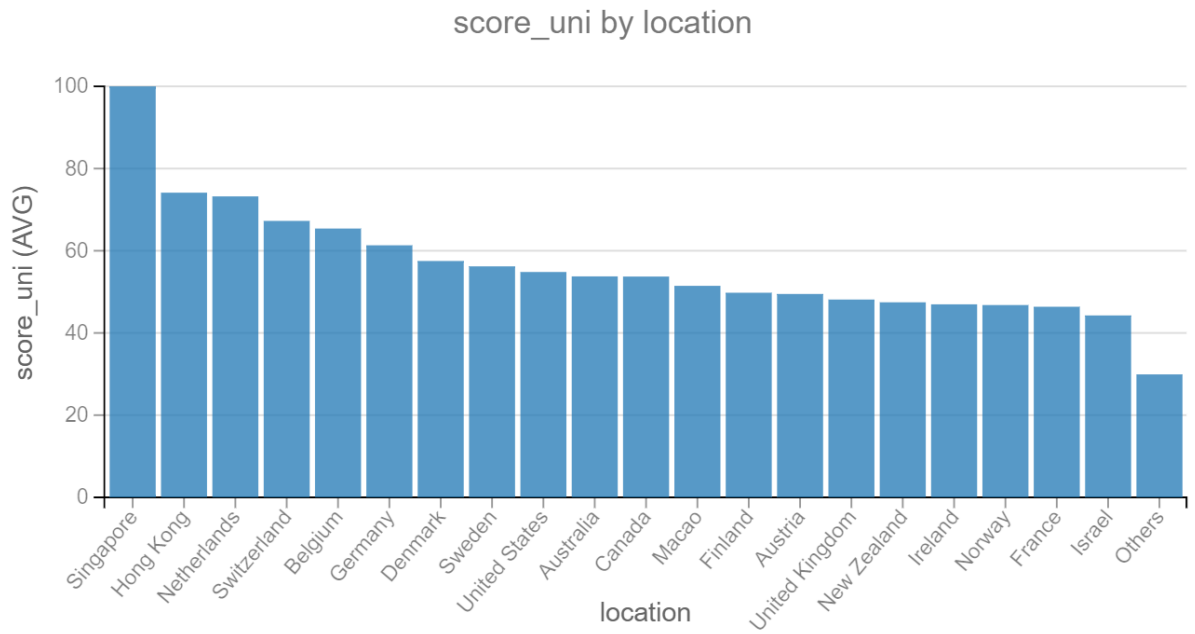


2. Analyse des universités :

Dans La deuxième partie nous allons étudier la qualité de l'enseignement supérieure de chaque pays.

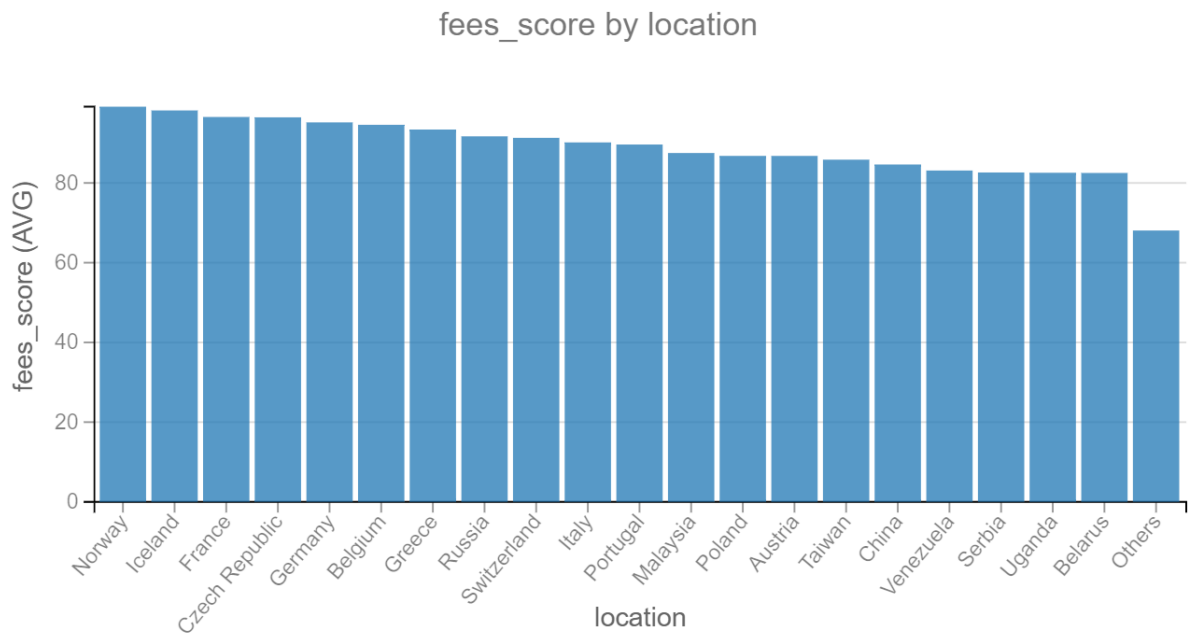
2.1. Qualité d'enseignement :

On commence par analyser la qualité de l'enseignement de chaque pays. Pour cela on définit un score qui nous permettons de faire cette étude (s'appeler score_uni dans notre modélisation) .On calcule cette métrique en utilisant la dimension score en faisant la moyenne pondéré sur le taux de recherche, taux d'insertion de chaque université et pour chaque pays.



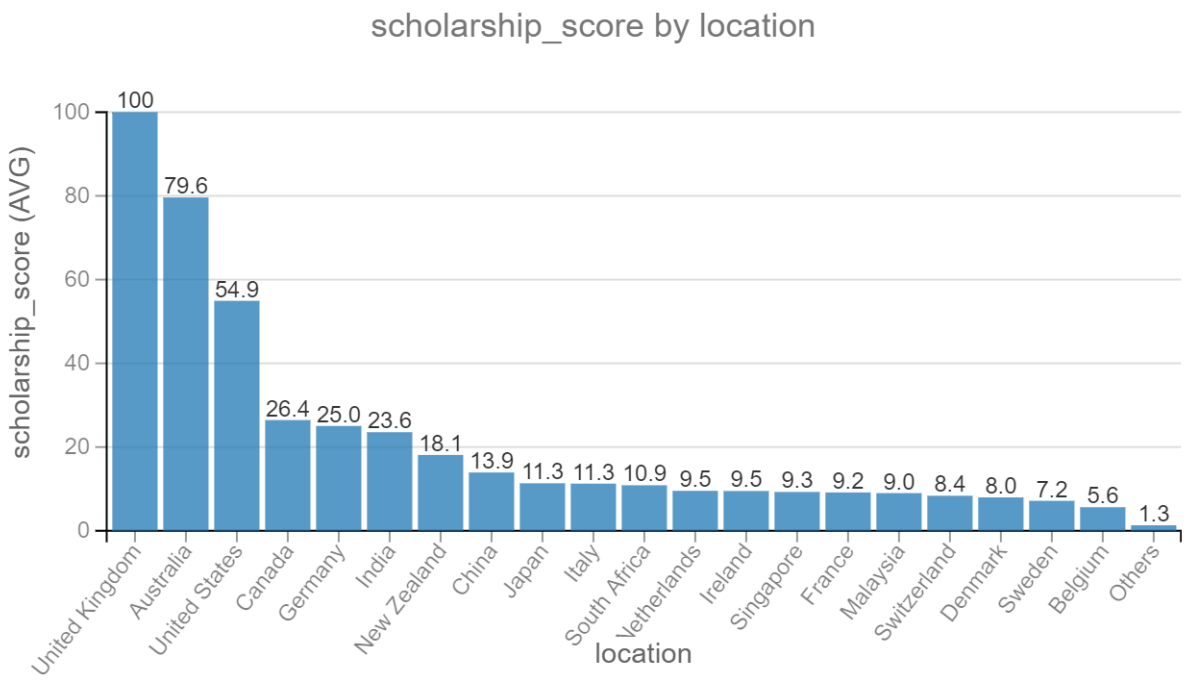
2.2. Frais d'inscription :

On fait aussi une étude sur les frais d'inscription en utilisant la dimension Fees et en faisant la moyenne des frais de licence et master pour chaque université. Puis on calcule la moyenne des frais pour chaque pays.

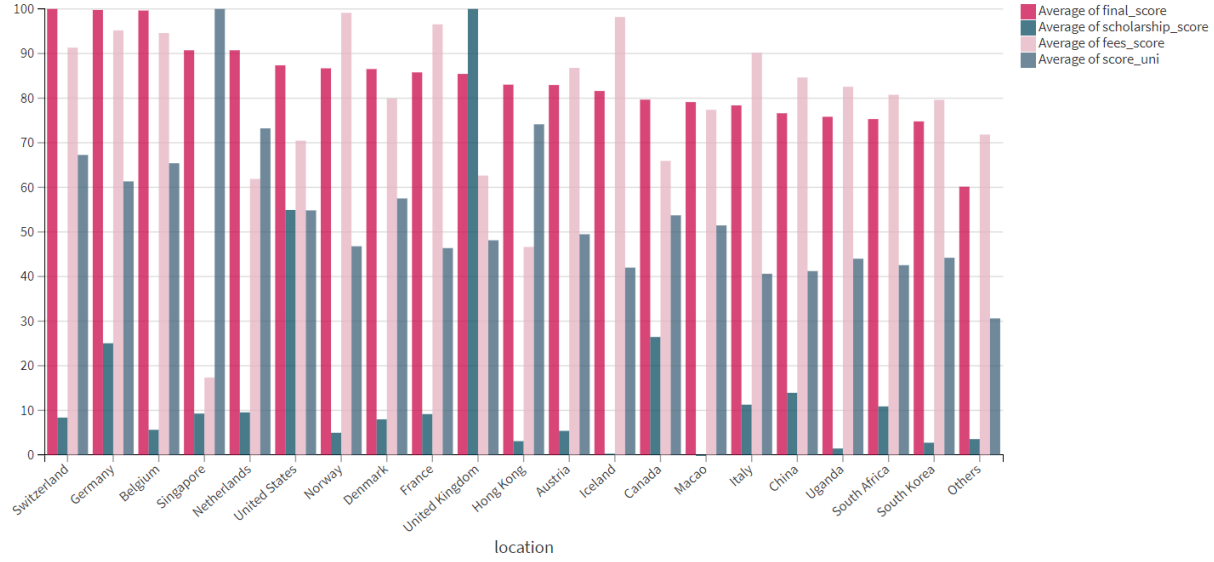


2.3. Bourses d'études :

Les bourses sont très importantes pour les étudiants et peuvent fortement affecter la décision de choix d'université. Donc on met aussi le nombre de bourse pour chaque pays comme mesure en utilisant la dimension Scholarship.

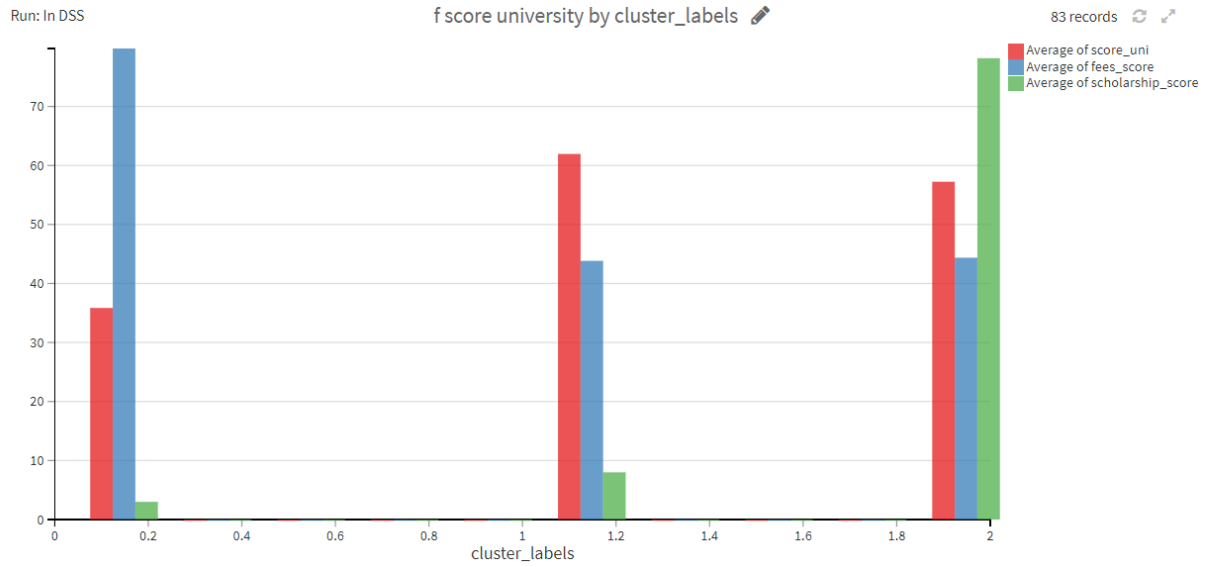


L'histogramme suivant contient une synthase sur les analyses des métriques précédents trié par leur score combiné.



2.4. Clustering :

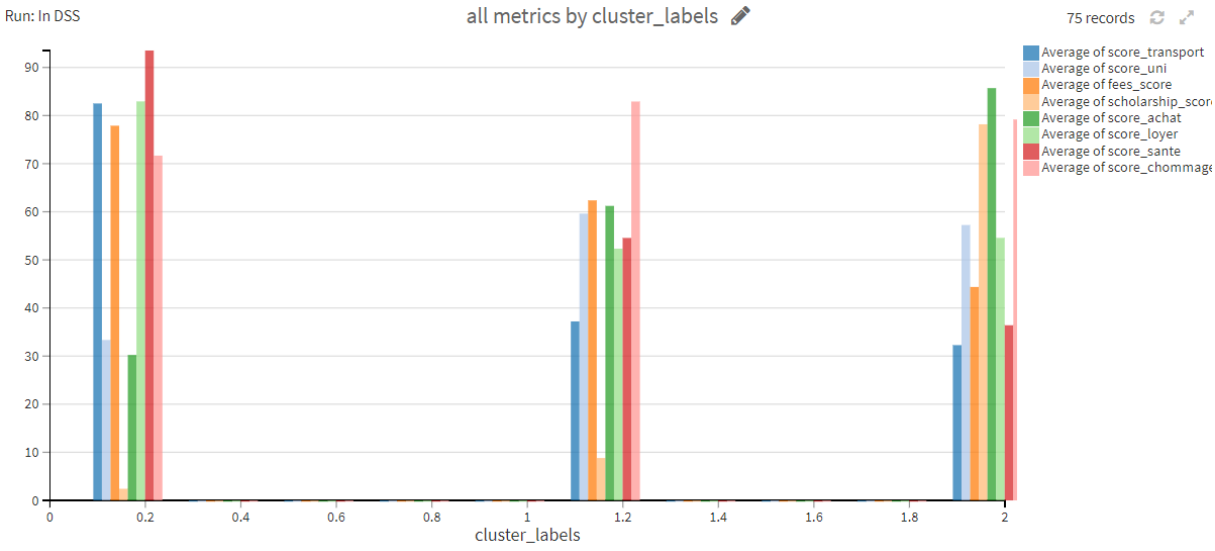
Nous avons fait un clustering (k-means) (avec un k=3) sur les 3 mesures décrite précédemment.



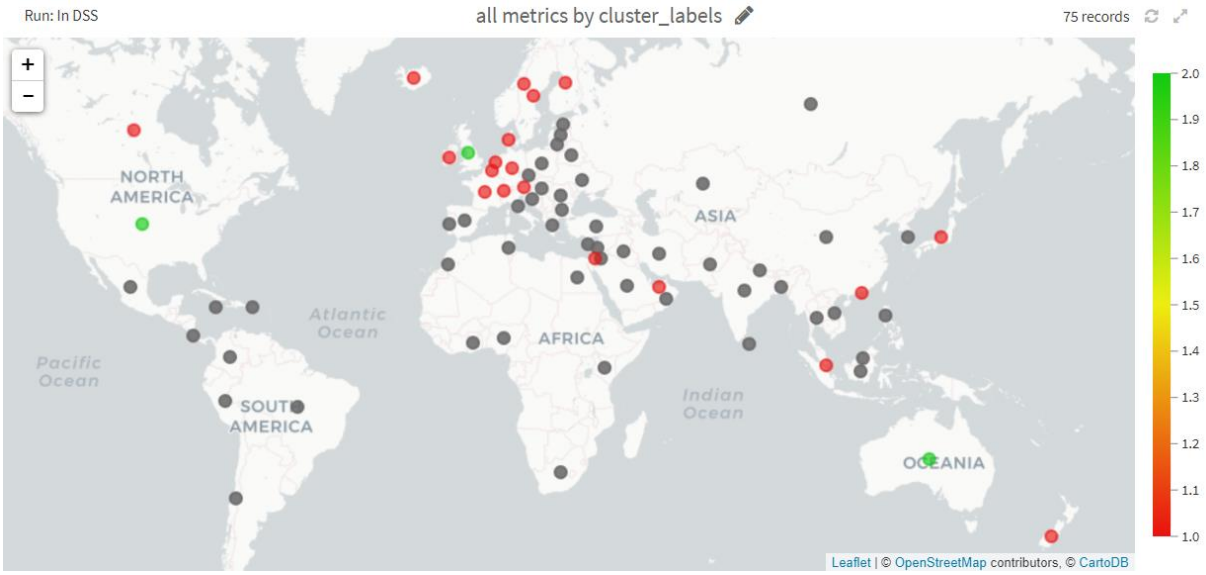
Nous pouvons identifier trois groupes de pays : le premier groupe de pays qui ne proposent pas trop de bourse mais leurs frais d'inscription sont petits. Un deuxième groupe de pays qui proposent plus de bourses mais les frais d'inscription sont plus élevés en même temps la qualité d'étude est bonne. Enfin, le troisième groupe qui est assez équilibrer, une bonne qualité d'enseignement, beaucoup de bourse et des frais d'inscription moyenne.

3. Résolution de la problématique :

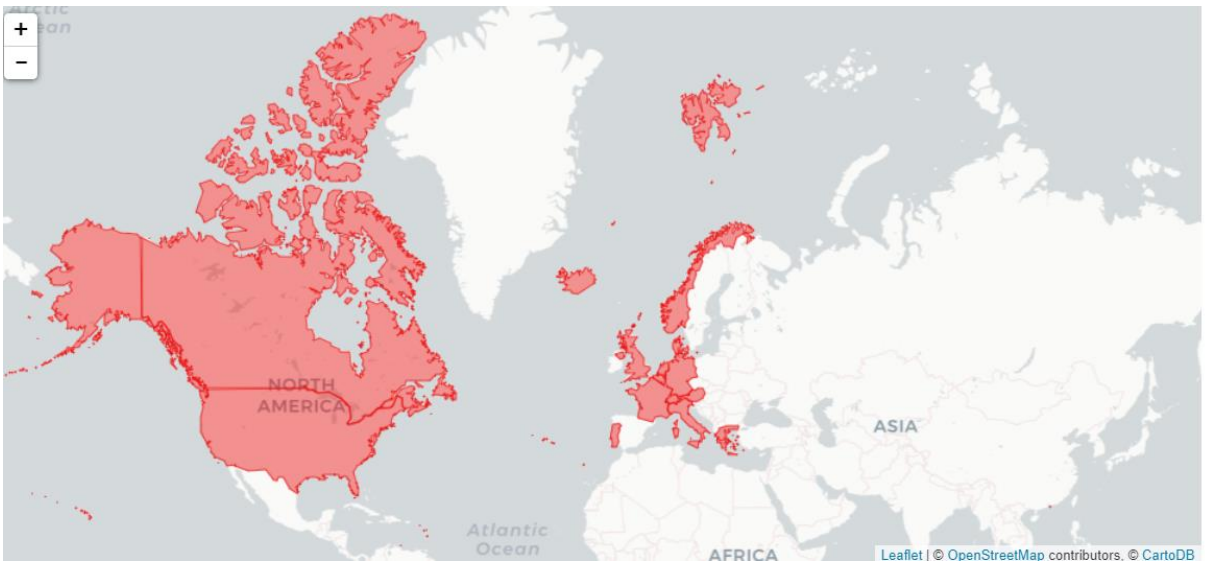
On regroupe les résultats des deux analyses et on fait un clustering sur les 9 métrique (score_trasport, score_uni, fees_score, scholarship_score, score_achat, score_loyer, score_sant, score,chommage),on trouve les résultats suivant :



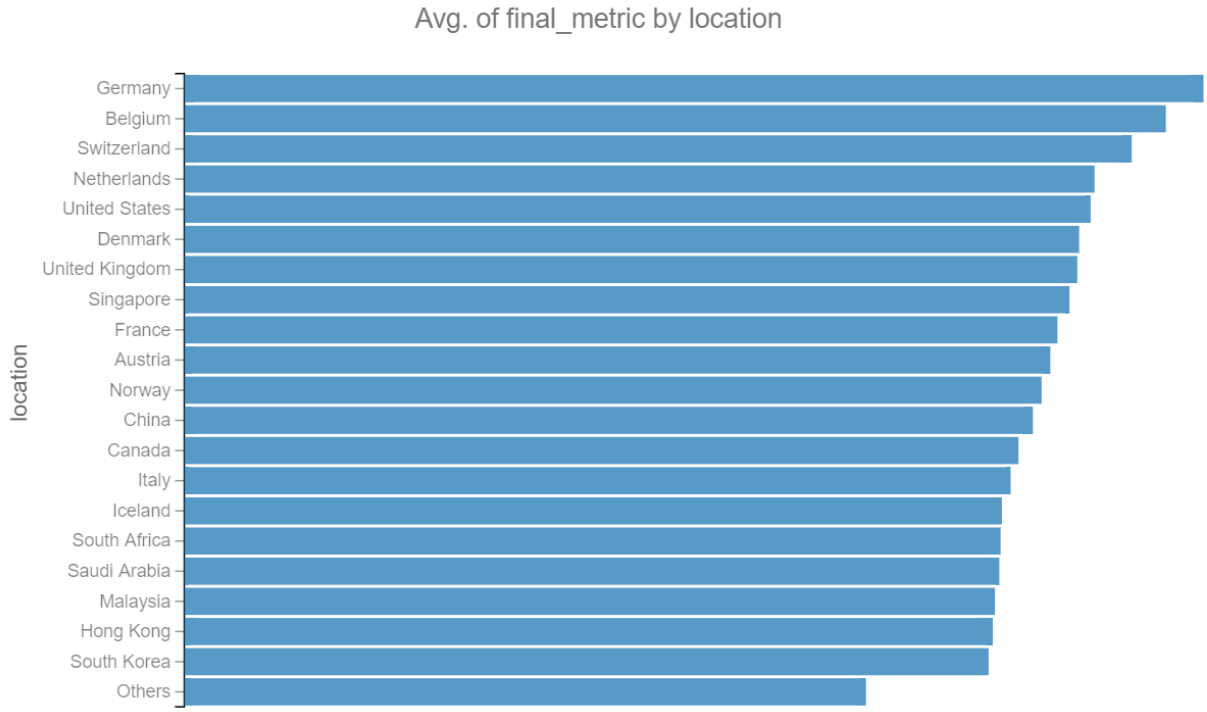
Sur la carte ci-dessus on affiche les pays par cluster : cluster_0 en gris, cluster_1 en rouge et cluster_2 en vert. Même si le cluster_0 a des bons scores au frais de transport, sante et inscription, le pouvoir d'achat et la qualité d'enseignement sont très faible, le taux de chômage est élevé et les bourses sont presque inexistant. Le cluster_1 semble plus équilibré malgré le manque des bourses les autres mesures sont acceptables. Enfin, Le cluster_2 semble aussi bien sauf que les frais de transport et santé sont plus élevé.



On a essayé de trouver aussi un score général pour les deux problèmes en utilisant une somme pondérée sur les métriques de chaque problème. Ensuite, on affiche sur la carte suivante les pays avec les meilleurs résultats :



On affiche aussi les pays triés par score, dans la première place c'est **l'Allemagne**.



V. Conclusion :

Dans un premier temps, nous avons analyser les métriques associer à l'état des pays en générale qui peut nous aider dans notre choix. Nous avons pu observer 3 groupes différents de pays ayant des caractéristiques propres sur le coût de vie. Ensuite, nous avons cherché à analyser la qualité de l'enseignement de chaque pays en prenant compte aussi les bourses d'études offertes et les frais d'inscription. On a pu identifier avec le clustering des paternes entre les pays pour les classifié en groupes. Enfin, en utilisant une combinaison des métriques nous avons trouvé les pays les plus appropriés pour continuer ses études supérieures.