Technique rapport BIUM

Haoran LI / Marcelin SEBLE

Contents

1 Problématique

Quel est l'État américain le plus approprié pour élever ses enfants?

2 Sources dataset

Wikipedia (Parce que nous faisons un dataset nous-même en utilisant Wiki) :
https://www.wikipedia.org/
Dataworld: https://data.world/
USnews: https://www.usnews.com/
Moneyrates: https://www.moneyrates.com/
Urban Institute Data Catalog: https://datacatalog.urban.org/
Statista : https://www.statista.com/

3 Datasets utilisés

1. State Population by Characteristics: 2010-2019
https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html
2.Crime 2017 per100k
https://worldpopulationreview.com/state-rankings/crime-rate-by-state
3. All data of GDP(2019-2020)
https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_GDP
4. health-care rank
https://www.usnews.com/news/best-states/rankings/health-care
5. State by State spending per child dataset (public health, librairies spending, housing and community development and parks expenditure) data from 1997 to 2016
https://datacatalog.urban.org/dataset/state-state-spending-kids-dataset
6. USA's cities demographic information (this dataset was created in 2015, last update in 2017)
https://public.opendatasoft.com/explore/dataset/us-cities-demographics/table
7. Median age by state
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_median_age 8. Crime rate
https://www.statista.com/statistics/301549/us-crimes-committed-state/
9. School-expenditure-by-state
https://www.statista.com/statistics/306693/us-per-pupil-public-school-expenditure-by-state
10. Higher-education-institutions-by-state
https://www.statista.com/statistics/306880/us-higher-education-institutions-by-state/
11. Education-expenditure-as-percent-of-government-expenditures
https://www.statista.com/statistics/306680/us-education-expenditure-as-percent-of-government-expenditures/
12. Per-capita-personal-income
https://www.statista.com/statistics/303555/us-per-capita-personal-income/
13. Closing-costs-by-state
https://www.statista.com/statistics/888157/closing-costs-by-state-usa/
14. Full-time-care-cost-in-family-care
https://www.statista.com/statistics/253998/full-time-care-cost-for-a-school-age-child-in-family-care-in-the-us-by-state/
15. Full-time-care-cost-in-a-child-care-center
https://www.statista.com/statistics/253938/full-time-care-cost-for-an-infant-in-a-child-care-center-in-the-us-by-state/

16. GDP by state
https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_GDP

4 Traitement des données

4.1 Normalisations des valeurs

Dans ce projet, on a rentré beaucoup de valeurs de différents types, quantités ou pourcentages. Pour les valeurs de sens positif, comme revenu personnel, le plus haut sera le mieux. On utilise $(MAX - X)/(MAX - MIN)$, donc la métrique est entre 0 et 1.
Par contre, pour les valeurs de sens négatif, comme taux de criminalité, le plus bas sera le mieux. On utilise $(X - MIN)/(MAX - MIN)$, la métrique est aussi entre 0 et 1.
Comme on a vu qu'il y a plusieurs dimensions pour cette problématique, on utilise la **moyenne pondérée** pour décrire le fait. Le score total doit être entre 0 et 1.

4.2 ETL

Ici on discute des manipulations de Dataiku qui s'appellent Visual recipes, "Prepare", "Join With"
• On supprime quelques lignes ou colonnes pas utiles. Normalement, les données sont avec différentes années, on choisit l'année la plus proche.
• En utilisant la méthode de moyenne pondérée, on a besoin de créer une nouvelle colonne avec une formule.
• Pour combiner plus d'une table, on fait la jointure avec la clé commune.
• Si on rencontre quelques données manquantes, on remplace les trous avec la moyenne. C'est une façon générale.

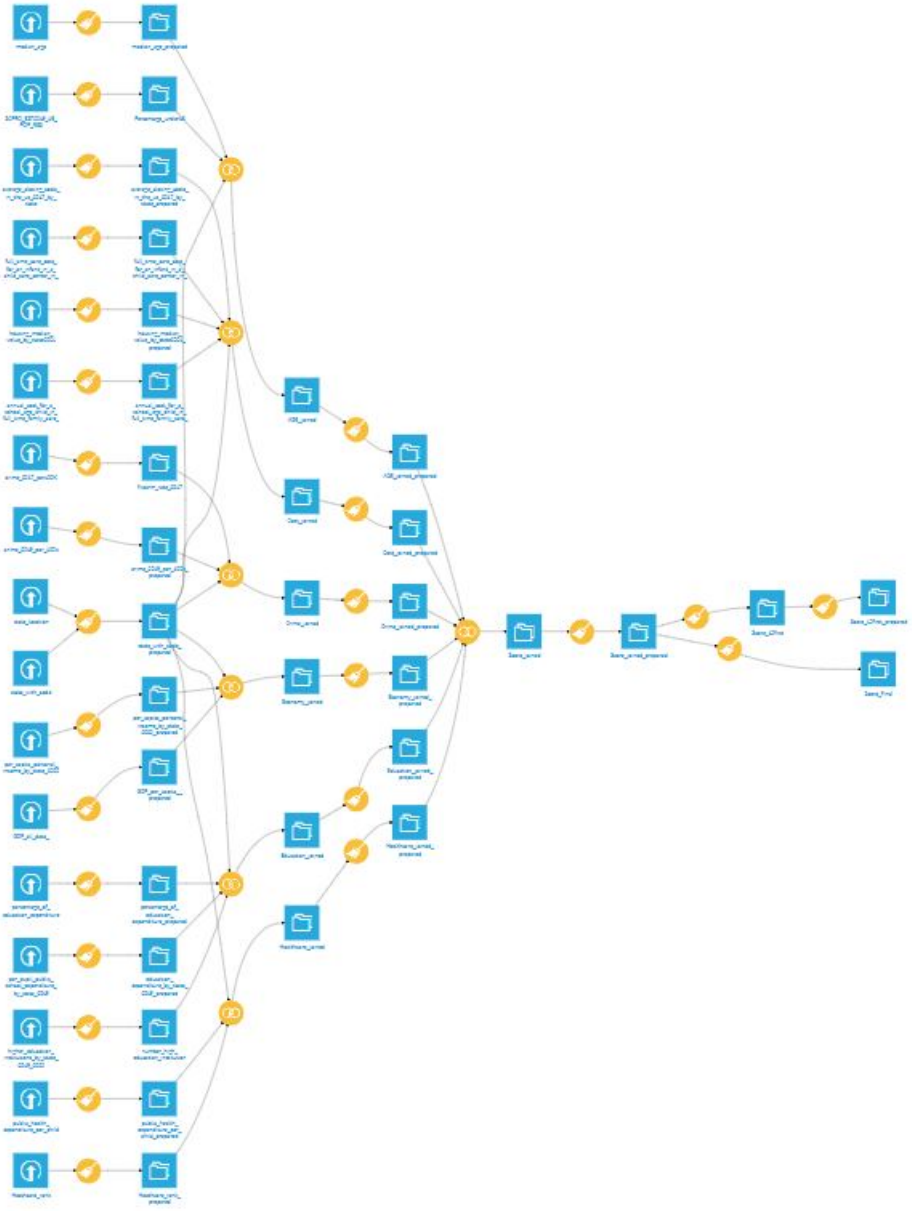

Figure 1: Dataiku Flow Screen Shot

5 Machine Learning

On fait sur les 6 dimensions : Crime score \ Cost score \ Education score \ Economy score \ Age score \ Healthcare score

5.1 PCA

PCA est une méthode linéaire de reduction de dimension. Par exemple, ici il y a 6 dimension pour nos données, si on veut les voir sur 2D espace. On doit les projeter sur les axes principaux commes ci-dessous:

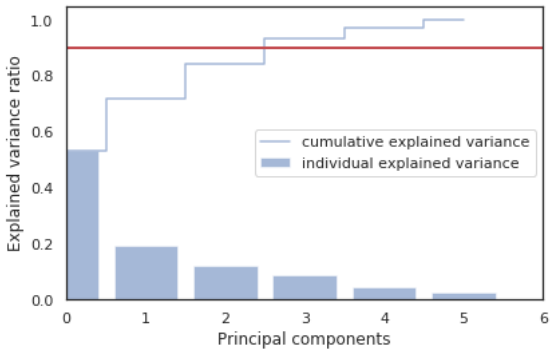

Figure 2: Cumulative Variance

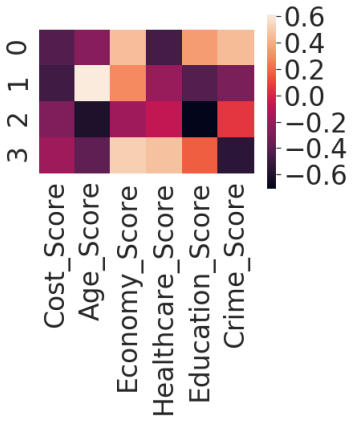

Figure 3: Heatmap

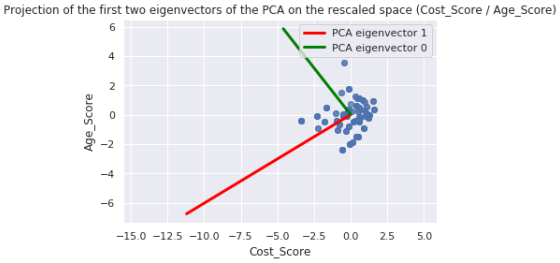

Figure 4: proj on the rescaled space

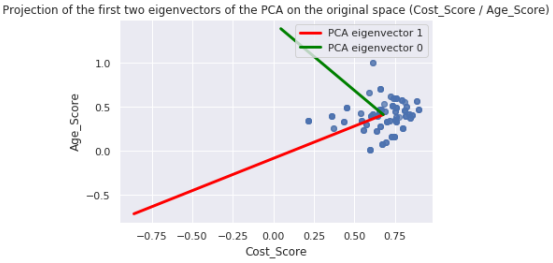

Figure 5: proj on the original space

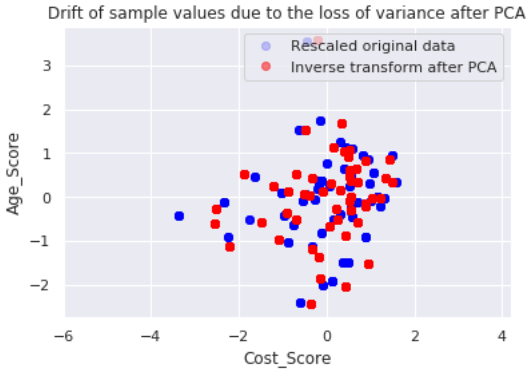

Figure 6: After PCA

5.2 T-sne

T-sne est une méthode non-linéaire de réduction. Elle est aussi une très bien visualisation. comme on voit ci-dessous, on peut changer la perplexité pour le voir. On sait les relations et similartés entre les

états par cette méthode.

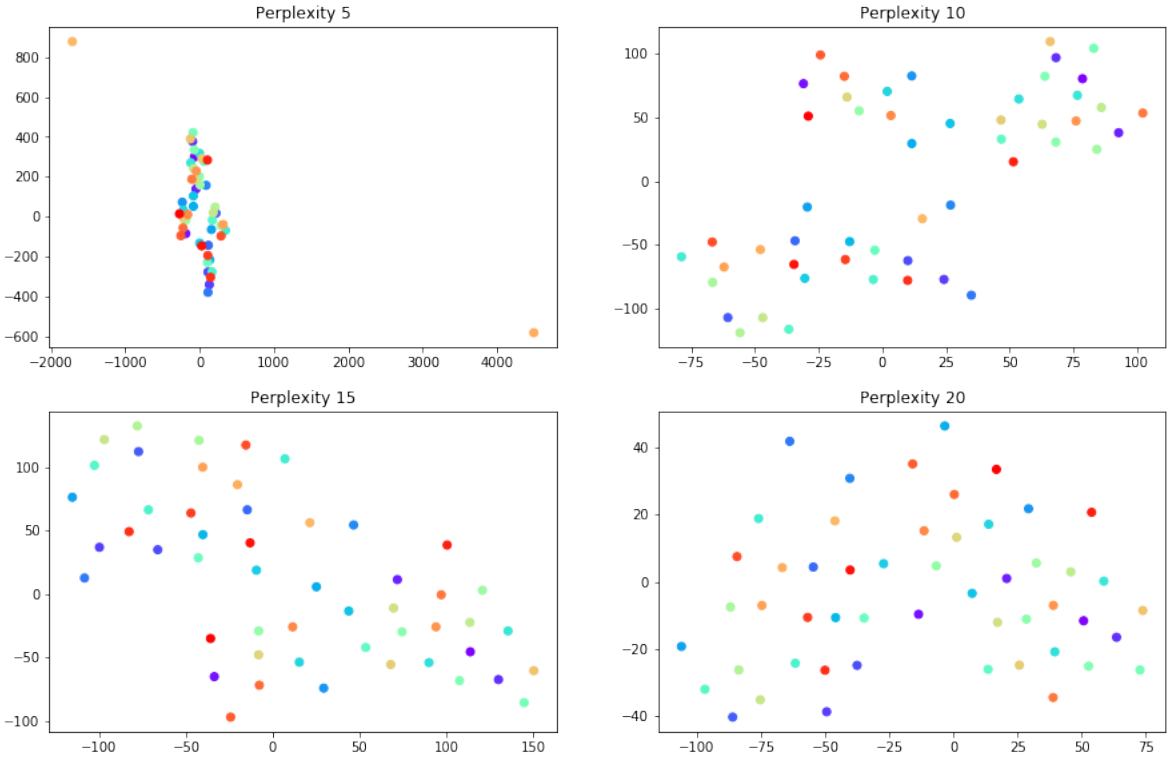

Figure 7: Visualisation T-sne

5.3 Clustering

On fait sur Dataiku, en utilisant différents modèles. En effet, c'est pas très utile dans notre projet, mais on les essaye.

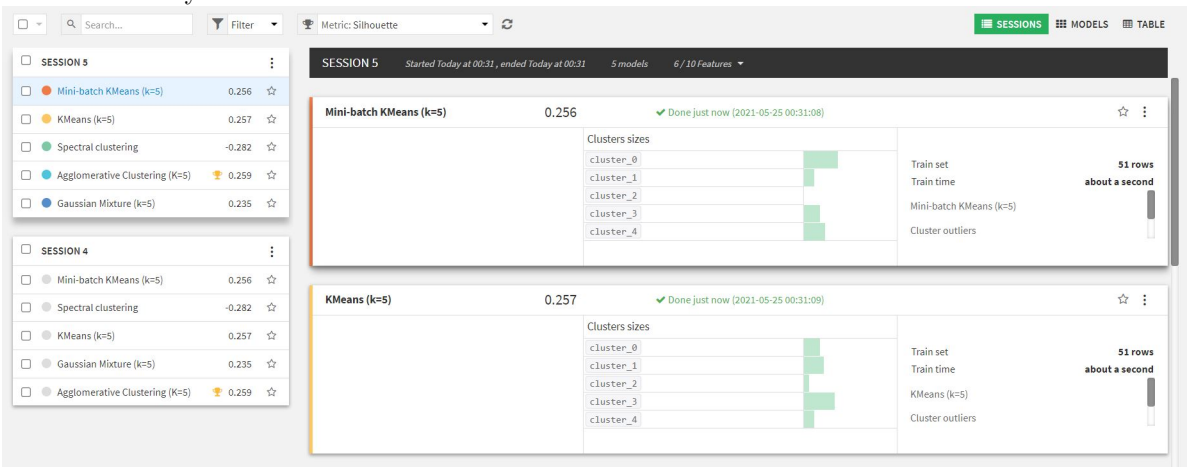

Figure 8: Clustering

Tous les choses qu'on fait en Dataiku, vous pouvez le consulter.

6 Vérification

On trouve 2 palmarès ci dessous:
https://www.moneyrates.com/research-center/10-best-states-for-raising-children.htm
https://www.usnews.com/news/best-states/articles/2021-01-11/massachusetts-is-best-state-to-raise-a-family-report-shows

Top Ten States to Raise a Family

1. New Jersey
2. (tie) Massachusetts
3. (tie) Utah
4. Connecticut
5. New Hampshire
6. Wyoming
7. (tie) Florida
8. (tie) Wisconsin
9. Ohio
10. New York

Figure 9: Palmarès 1 sur Internet

Best States to Raise a Family

Here are the 10 best states to raise a family, according to WalletHub:

1. Massachusetts
2. Minnesota
3. North Dakota
4. New York
5. Vermont
6. New Hampshire
7. New Jersey
8. Washington
9. Connecticut
10. Utah

Figure 10: Palmarès 2 sur Internet

On fait un palmarès en utilisant notre modèle comme ci-dessus.

State_name	Total_Score
string US State	double Decimal
New Jersey	0.5710937690895755
Utah	0.5449582451094553
Wyoming	0.5406797338380177
Connecticut	0.5386093364641973
New York	0.5297231964248368
North Dakota	0.5260675744640979
Nebraska	0.5253599929138123
Texas	0.5223882888151133
Massachusetts	0.5169924669027175
Kansas	0.5119218262095991
South Dakota	0.5019909461650717
Alaska	0.49131815483727953
Iowa	0.49070545203766675
Georgia	0.4891039640379269
New Hampshire	0.48666432842171187

Figure 11: Notre palmarès

Ils sont en bon accord. Cela montre que notre travail est assez crédible.

• Remarque:
Si tu veux consulter tous les choses qu'on montre(les tables, les traitements, les ML), je vous invite d'aller dans notre Dataiku.
Username :lih / Password : haoran