Sistemas Inteligentes

Projeto da Disciplina
Etapa 2

Alunos:

Bruno Henrique Da Silva Lucena

João Victor Voltarelli

Rodrigo Leonello Bellotti

- ## Descrição da base de dados

  A base de dados utilizada contém dados sobre o COVID-19 separados por estados dos EUA, juntamente com os dados da parte econômica de cada estado.

  A base está disponível em: https://www.kaggle.com/nightranger77/covid19-state-data

- ## Objetivo

  O objetivo principal é, com base nessa base de dados, descobrir se a situação econômica de um determinado estado afeta mais ou menos a transmissão, casos positivos e negativos e mortes relacionados ao vírus do COVID-19.

- ## Pré-processamento dos dados

  Primeiramente, no pré-processamento dos dados, nós analisamos todos os atributos e descartamos os que julgamos não serem inicialmente importantes para o nosso modelo. Com isso, mantivemos na base somente os atributos diretamente relacionados a saúde e aos dados econômicos:

  State – Indica o nome do estado.

  Tested – Indica o número de pessoas testadas para o COVID-19.

  Infected – Indica o número de pessoas infectadas com o COVID-19.

  Deaths – Indica o número de pessoas mortas devido ao COVID-19.

  Population – Indica o número da população do estado.

  Pop Density – Indica a densidade populacional do estado.

  Gini – Indica o índice de desigualdade do estado (Coeficiente de Gini).

  ICU Beds – Indica o número de leitos de UTI.

  Income – Indica a receita/renda do estado.

  GDP – Indica o produto interno bruto do estado (PIB).

  Unemployment – Indica o número de desempregos do estado.

  Hospitals – Indica o número de hospitais do estado.

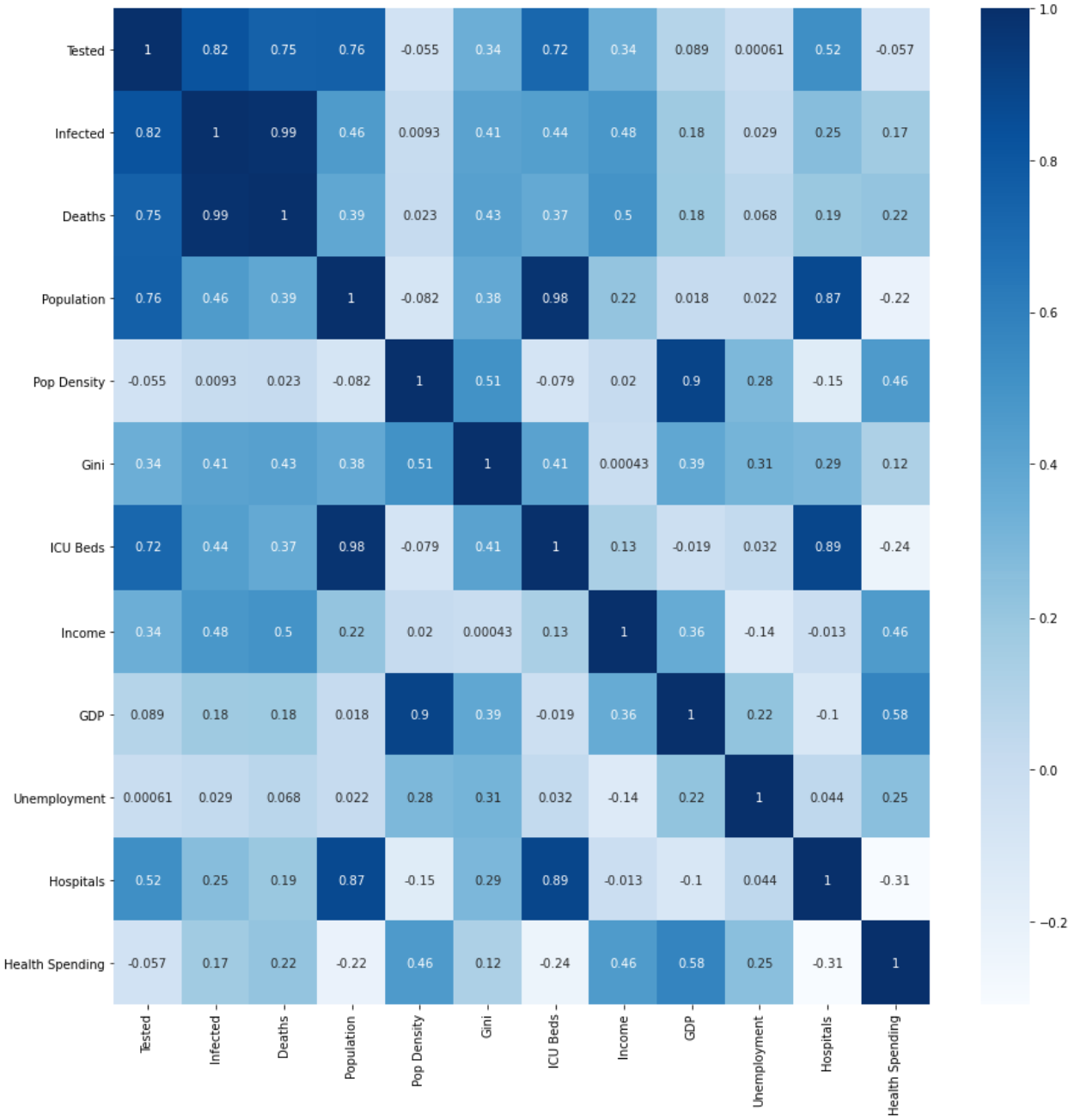
  Health Spending – Indica a quantidade de gastos com a área da saúde.

Após essa análise inicial, utilizamos alguns algoritmos em Python para nos auxiliar no pré-processamento dos dados. Através destes algoritmos conseguimos gerar uma matriz de correlação e um mapa de calor, utilizados para descobrirmos quais as relações que os atributos têm entre si.

- **Matriz de correlação gerada**

	Tested	Infected	Deaths	Population	Pop Density	Gini	ICU Beds	Income	GDP	Unemployment	Hospitals	Health Spending
Tested	1.000000	0.824178	0.753604	0.759423	-0.054662	0.343993	0.719167	0.344514	0.088660	0.000609	0.520018	-0.057032
Infected	0.824178	1.000000	0.985206	0.456149	0.009313	0.411899	0.435392	0.483377	0.176994	0.028866	0.253490	0.168867
Deaths	0.753604	0.985206	1.000000	0.386997	0.022900	0.427113	0.371621	0.502059	0.183368	0.068290	0.194121	0.215194
Population	0.759423	0.456149	0.386997	1.000000	-0.082282	0.380073	0.978022	0.216398	0.018147	0.021802	0.873197	-0.219486
Pop Density	-0.054662	0.009313	0.022900	-0.082282	1.000000	0.506948	-0.079125	0.019956	0.898326	0.284099	-0.152769	0.460614
Gini	0.343993	0.411899	0.427113	0.380073	0.506948	1.000000	0.414981	0.000428	0.390454	0.311418	0.291098	0.123952
ICU Beds	0.719167	0.435392	0.371621	0.978022	-0.079125	0.414981	1.000000	0.132944	-0.018564	0.031786	0.889141	-0.235379
Income	0.344514	0.483377	0.502059	0.216398	0.019956	0.000428	0.132944	1.000000	0.362317	-0.138739	-0.013090	0.456135
GDP	0.088660	0.176994	0.183368	0.018147	0.898326	0.390454	-0.018564	0.362317	1.000000	0.217973	-0.104310	0.580269
Unemployment	0.000609	0.028866	0.068290	0.021802	0.284099	0.311418	0.031786	-0.138739	0.217973	1.000000	0.044494	0.247427
Hospitals	0.520018	0.253490	0.194121	0.873197	-0.152769	0.291098	0.889141	-0.013090	-0.104310	0.044494	1.000000	-0.309285
Health Spending	-0.057032	0.168867	0.215194	-0.219486	0.460614	0.123952	-0.235379	0.456135	0.580269	0.247427	-0.309285	1.000000

- **Mapa de calor gerado**



Com uma análise inicial, através do mapa de calor, foi possível observar algumas informações interessantes, como:

- O atributo *Income* que seria a renda/receita de um estado, tem uma relação interessante com o atributo *Deaths*, que indica a quantidade de mortes causadas pelo COVID-19. Assim como, com o atributo *Infected*, que indica a quantidade de infectados pelo vírus.

- É interessante também observar o atributo *Gini*, que simboliza o índice de desigualdade de um estado. Ele também apresenta uma relação considerável com os atributos que indicam o número de infectados e mortos.

- O atributo GDP, que simboliza o PIB de um determinado estado, tem obviamente uma relação com o atributo *Health Spending*, que indica o gasto com a área da saúde, visto que, na teoria, quanto maior a renda interna de um estado, maior será o investimento nas diversas áreas, como a da saúde.

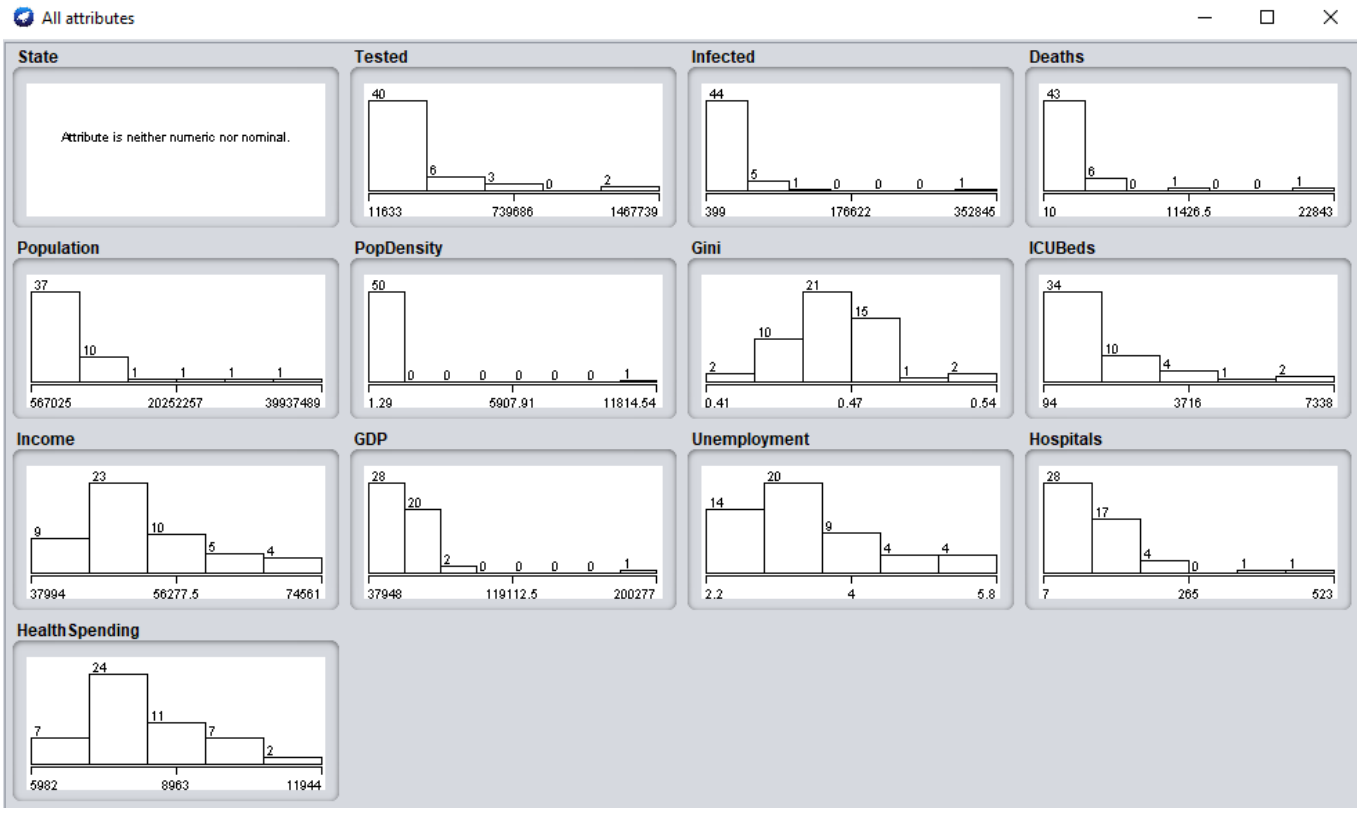
Essas análises iniciais, feitas com base na matriz de correlação e no mapa de calor, são essenciais e serão muito importantes para o desenvolvimento das atividades futuras, pois com isso, teremos de forma mais clara as relações dos atributos que serão necessários na realização do objetivo final.

- **Testes no Weka**

  Logo após esse pré-processamento inicial dos dados utilizando Python, utilizamos o Weka para observar os dados e rodar o algoritmo de clusterização K-means, para observar qual seria o comportamento. Obtivemos os seguintes resultados primários:

- **Visualização dos atributos**

  - Contendo o valor mínimo e máximo de cada estatística, e a quantidade de instâncias em cada faixa de valores.

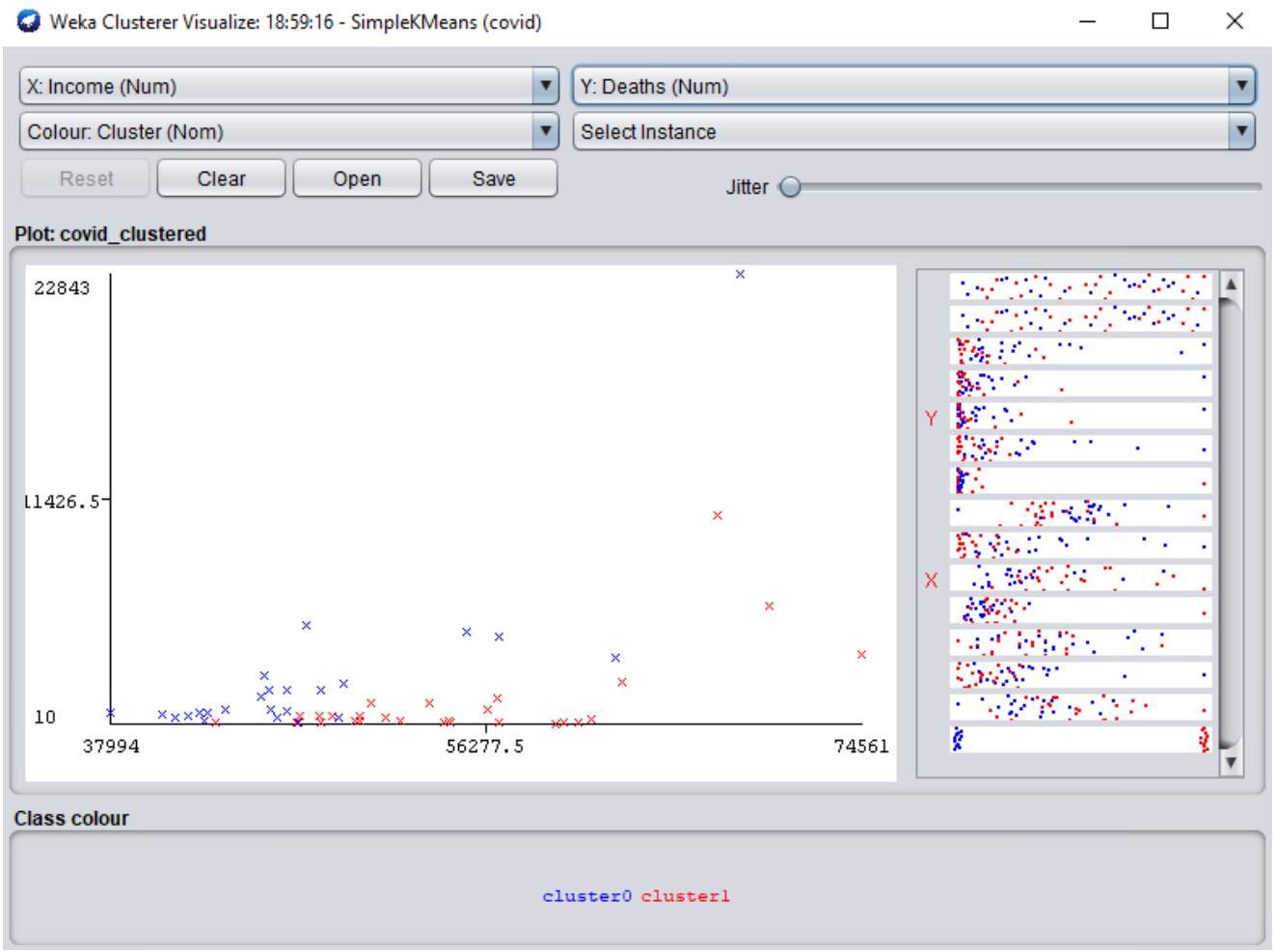


- **Testes com o algoritmo K-means**

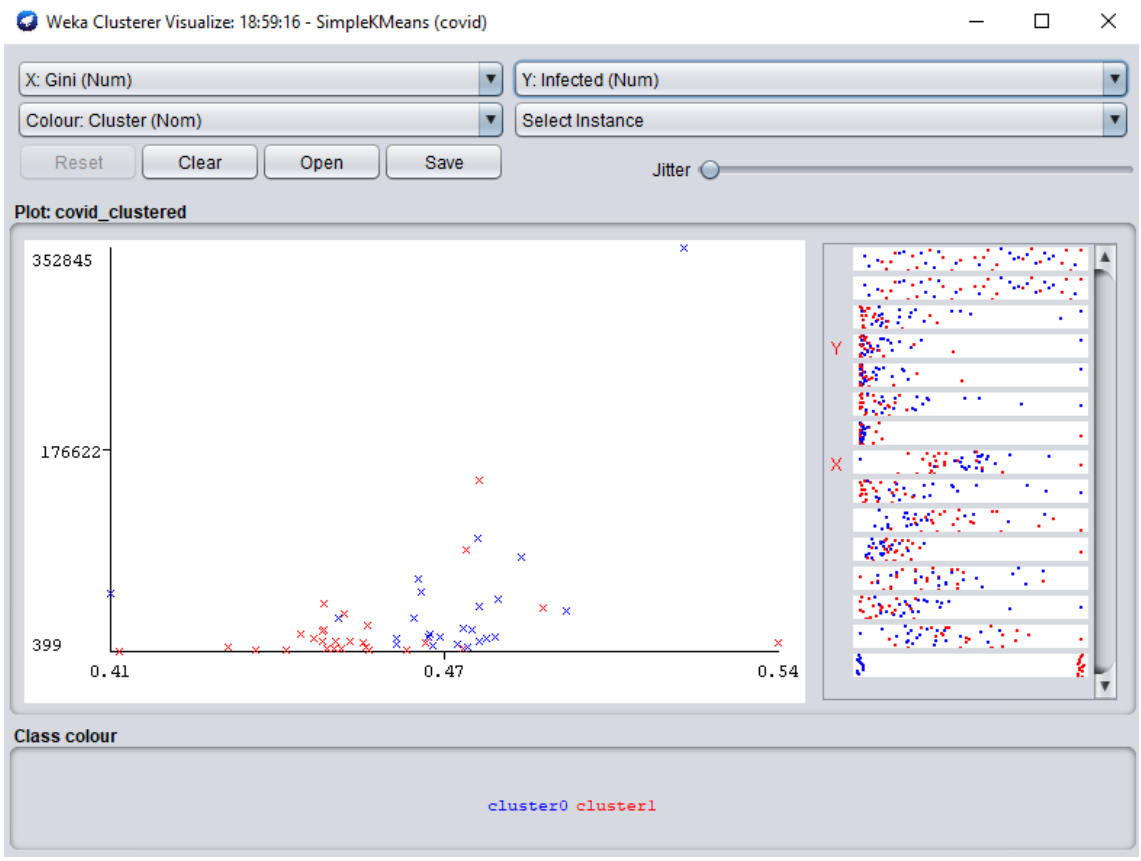
  Obs.: Não alteramos nenhum parâmetro utilizado no algoritmo. Foram utilizados os valores padrões do Weka para se observar como seria o comportamento inicial sem nenhuma alteração.

  Após rodarmos o algoritmo k-means com a base de dados, conseguimos visualizar alguns clusters gerados inicialmente. Abaixo estão três exemplos:

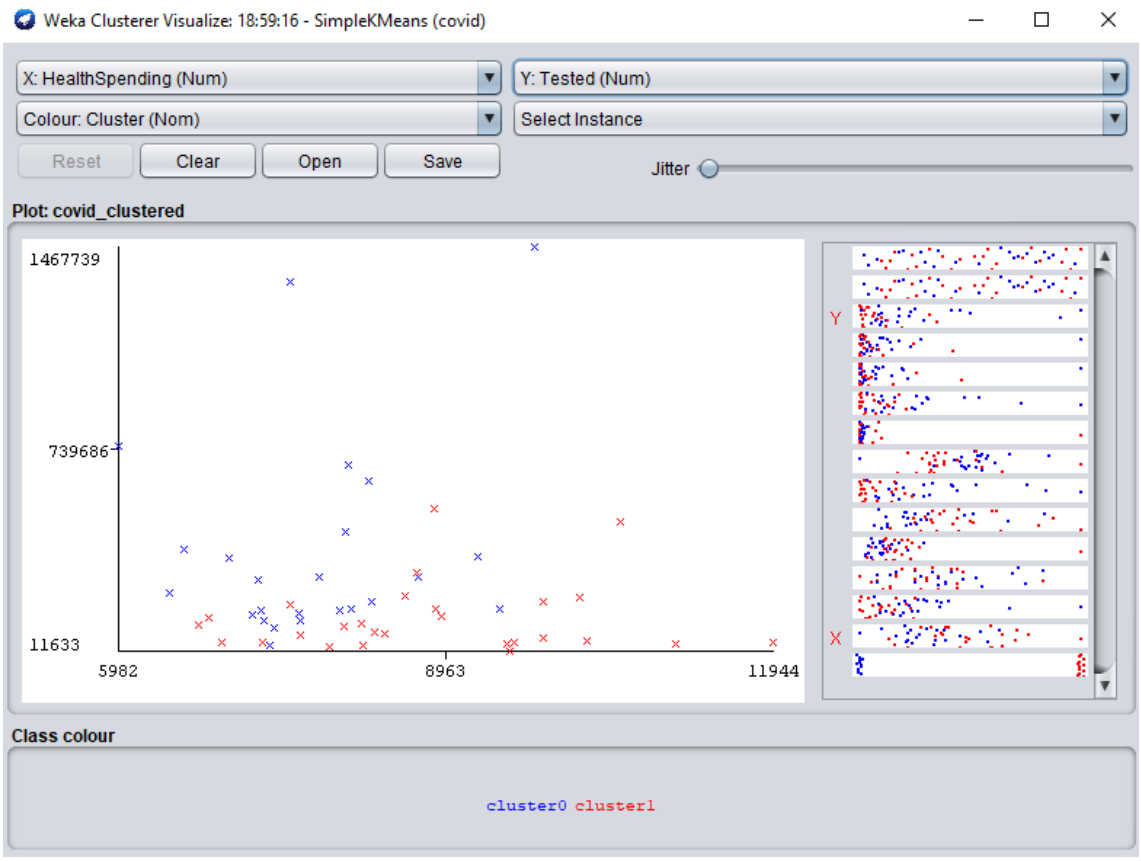
  - **X**: Income, **Y**: Deaths



- **X**: Gini, **Y**: Infected



- **X**: Health Spending, **Y**: Tested



- **Atividades futuras**

  Com base nos testes e resultados obtidos, continuaremos a analisar e ajustar os dados para que assim consigamos chegar no objetivo final, concluindo se há uma influência da economia de um estado no número de pessoas afetados pelo vírus da COVID-19 ou não.

- **Referências**

https://medium.com/@lucasoliveiras/primeiros-passos-com-kaggle-3871997b0868

https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

https://medium.com/@masonrchildress/how-to-make-a-correlation-heatmap-in-python-cc47e1c2fdc2

https://paulovasconcellos.com.br/como-selecionar-as-melhores-features-para-seu-modelo-de-machine-learning-2e9df83d062a

https://minerandodados.com.br/aprenda-como-selecionar-features-para-seu-modelo-de-machine-learning/