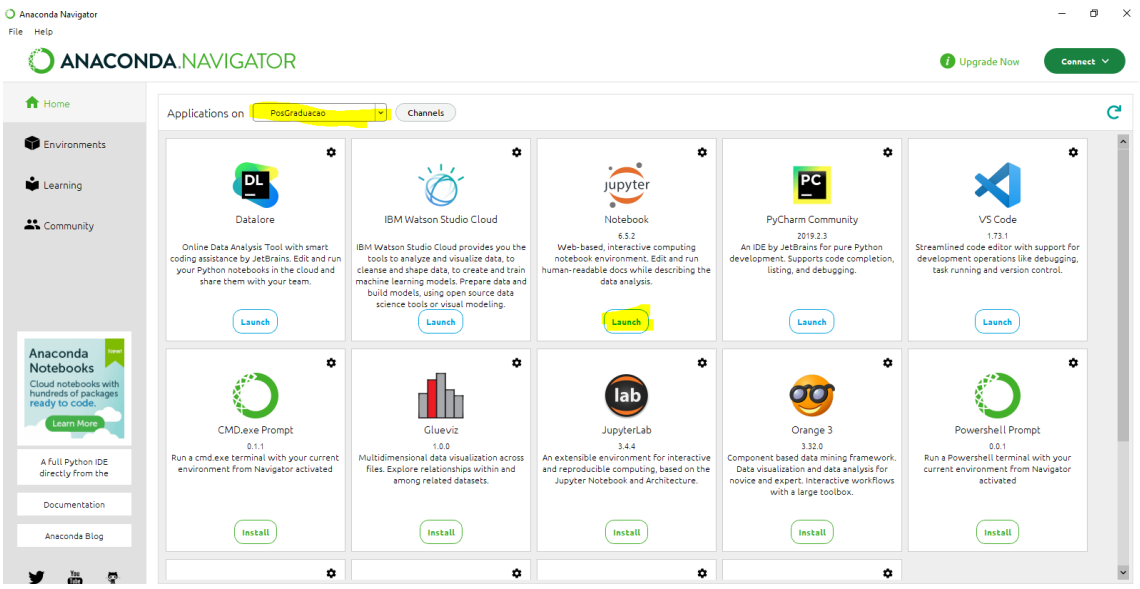
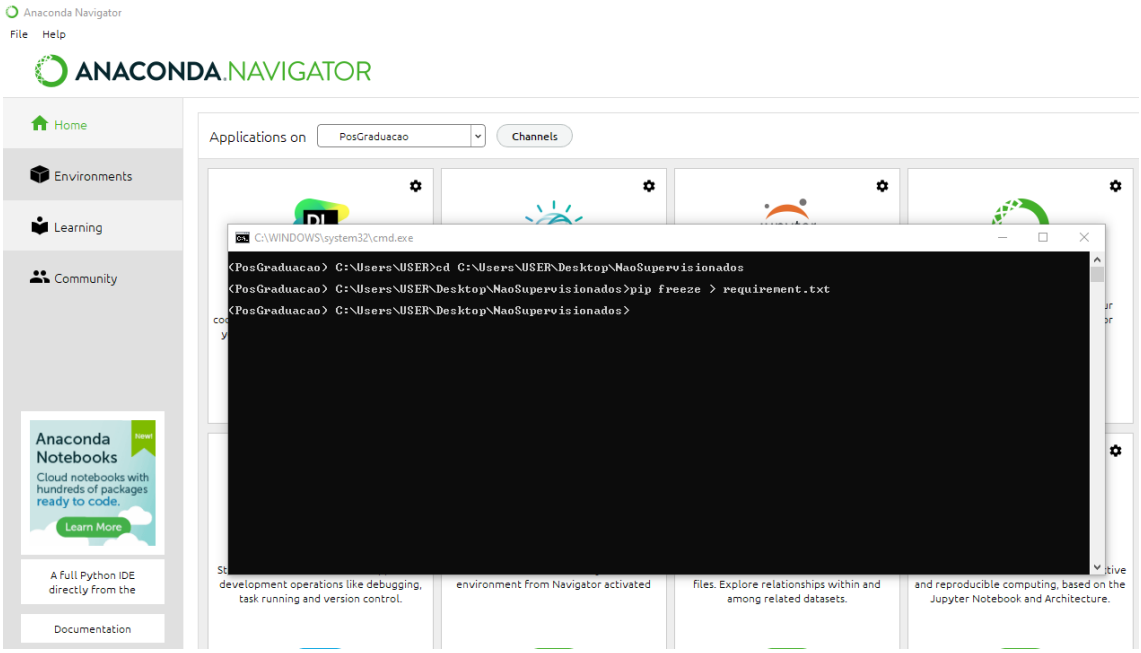
Respostas Teóricas Algoritmos Não-Supervisionados para Clusterização

Infraestrutura:

1 - Printscreen do ambiente:





Escolha de base de dados:

2 - Quantos países existem no dataset?

R: Existem 167 países no dataset.

3 - O que deve ser feito com os dados antes da etapa de clusterização?

R: A análise exploratória dos dados e realizar o pré-processamento para deixar os dados em formatos mais proveitosos e satisfatórios.

Essa é uma etapa crucial, uma vez que se o conjunto de dados não está preparado, não conseguiremos alcançar bons resultados, tornando as análises e modelagens não tão confiáveis.

Clusterização:

2 – Para os resultados, do K-Médias:

R:

A – Os clusters foram divididos com as seguintes dimensionalidade dos dados do DataSet:

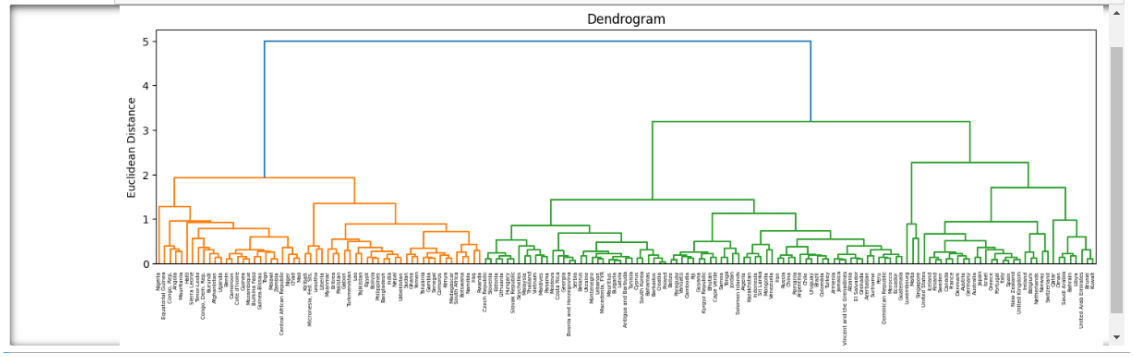
0 = 91 Países
1 = 44 Países
2 = 32 Países

3 –

Resultado da Clusterização Hierárquica

```
In [61]: df = paises.set_index('country')

In [72]: plt.figure(figsize=(16, 4))
         plt.grid(False)
         dendrogram = sch.dendrogram(sch.linkage(normalizados, method='ward'), labels=df.index)
         plt.title('Dendrogram')
         plt.ylabel('Euclidean Distance')
```



No cluster 0 se encontram os dados dos países com os índices socioeconômicos mais baixos enquanto o cluster 1 possuem índices um pouco melhores e o cluster 2 são os países com melhores índices.

4 –
R: Na Clusterização com o Kmeans, foi separado em 3 grupos:

No cluster 0 foram agrupados os países com o menor GDPP e maiores índices de mortalidade infantil, com o cluster 1 possuindo valores medianos desses índices e o grupo 2 possuem o maior GDPP e menor taxas de mortalidade infantil.

No dendrograma, a semelhança do Kmeans, foram organizados no 0 os países com maior índice de mortalidade infantil enquanto no cluster 2 estão agrupados os que possuem os maiores valores e no cluster 1 estão agrupados os países com os valores medianos.

Escolha de algoritmos:

1 - Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.

R:

A – Inicializar K centroides em pontos aleatórios
B – Para cada ponto, encontra qual o centroide mais próximo.
C – Calcular o baricentro dos pontos para cada centróide.
D – Mover o centróide na direção do seu baricentro.
E – Repetir a partir do tópico B.

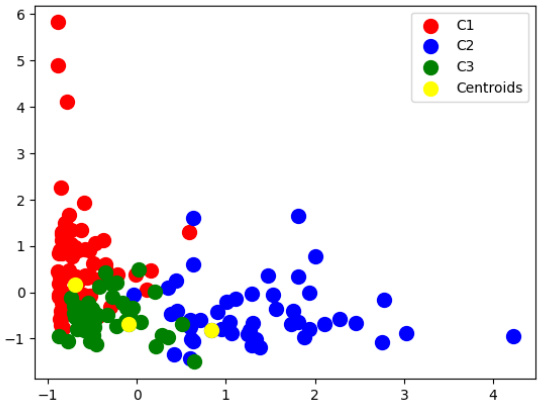
2 - O algoritmo de K-médias converge até encontrar os centróides que melhor descrevem os clusters encontrados (até o deslocamento entre as interações dos centróides ser mínimo). Lembrando que o centróide é o baricentro do cluster em questão e não representa, em via de regra, um dado existente na base. Refaça o algoritmo apresentado na questão 1 a fim de garantir que o cluster seja representado pelo dado mais próximo ao seu baricentro em todas as iterações do algoritmo.

Obs: nesse novo algoritmo, o dado escolhido será chamado medóide.

R: Algoritmo refeito com Kmedoides.

```
In [47]: plt.scatter(x_scaled[y_kmed == 0, 0], x_scaled[y_kmed == 0, 1], s = 100, c = 'red', label = 'C1')
         plt.scatter(x_scaled[y_kmed == 1, 0], x_scaled[y_kmed == 1, 1], s = 100, c = 'blue', label = 'C2')
         plt.scatter(x_scaled[y_kmed == 2, 0], x_scaled[y_kmed == 2, 1], s = 100, c = 'green', label = 'C3')
         plt.scatter(kmedoids.cluster_centers_[:, 0], kmedoids.cluster_centers_[:,1], s = 100, c = 'yellow', label = 'Centroids')
         plt.legend()

Out[47]: <matplotlib.legend.Legend at 0x2242b833f10>
```



3 - O algoritmo de K-médias é sensível a outliers nos dados. Explique.

R: O K-médias trabalha a base das medias do DataSet. Mas os Outliers, por se tratar de dados muito destoantes dos demais acaba modificando a média do DataSet.

Isso pode acarretar que no deslocamento dos centroides do seu local razoável. Logo, eles podem ficar longe dos lugares onde ele deveria estar efetivamente localizado.

4 - Por que o algoritmo de DBScan é mais robusto à presença de outliers?

R: O DBScan atua baseado em densidades e permite identificar grupos de diferentes formatos topológicos e identificar os ruídos isolados.