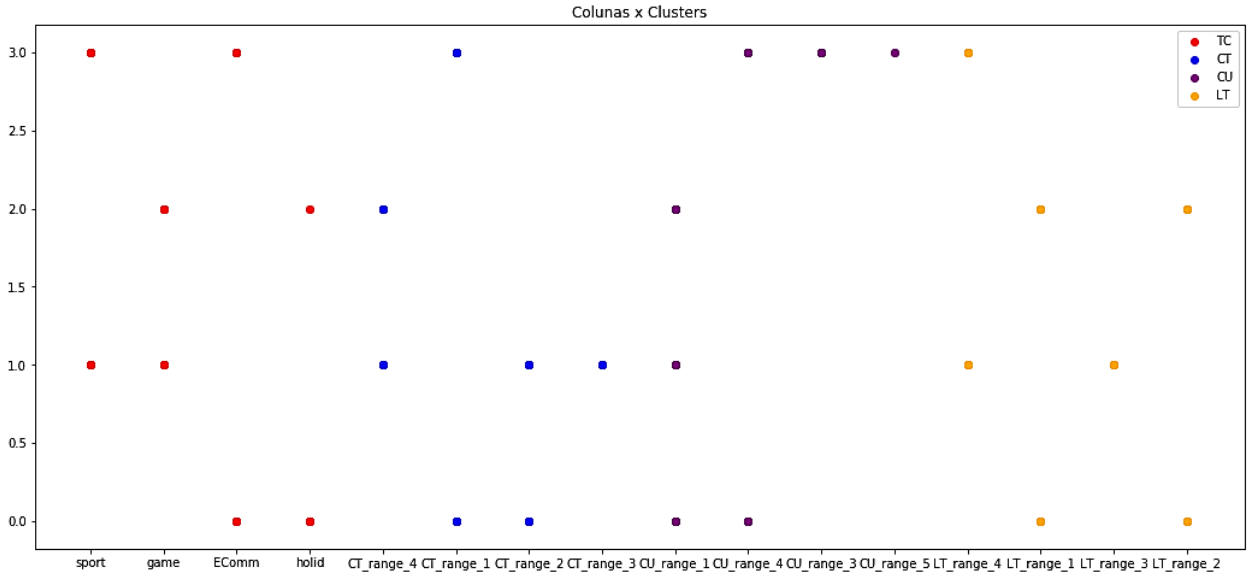
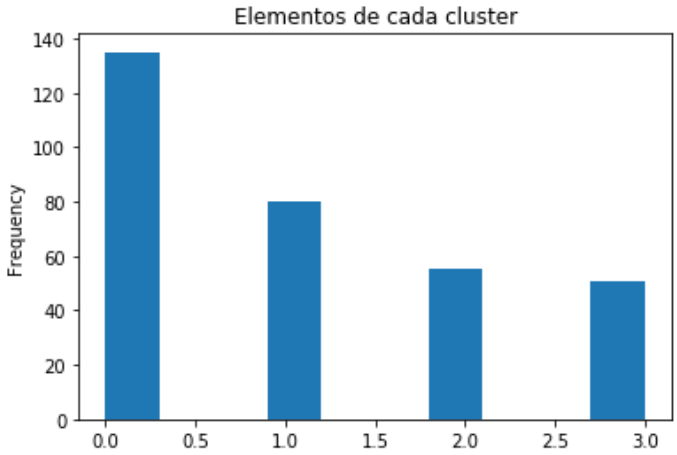
```
# -*- coding: utf-8 -*-
"""
@author: Laura Castro
"""

import numpy as np import pandas as pd
import matplotlib.pyplot as plt from
sklearn.cluster import KMeans from
sklearn.preprocessing import LabelEncoder

#ler o arquivo, e considerar as separações por ponto e
vírgula base_dados = pd.read_csv('bd_tarefa1.CSV', sep=';')
estatisticas = base_dados.describe()

"""PRÉ PROCESSAMENTO"""
#mudar os titulos das colunas (de acordo com o que ele nomeou na tarefa)
base_dados.columns = ['CT', 'CU', 'LT', 'TC', 'TS']


#é necessario ser vetor para conseguir clusterizar
#file.drop para retirar a coluna TS // axis=1 indica coluna, e axis=0
#indica linha
vetor = np.array(base_dados.drop(['TS'], axis=1))

#converter dados textuais em numéricos
#para interpretação do programa
vetor[:, 0] = LabelEncoder().fit_transform(vetor[:,0])
vetor[:, 1] = LabelEncoder().fit_transform(vetor[:,1])
vetor[:, 2] = LabelEncoder().fit_transform(vetor[:,2])
vetor[:, 3] = LabelEncoder().fit_transform(vetor[:,2])


"""CLUSTERIZANDO"""
clusterizar = KMeans(n_clusters=4, random_state=0)
#variável que vai guardar a clusterizacao
#n_clusters = define o numero de "clusters" (grupos) - como são 4 colunas
#random_state = numero de vezes que se inicia os clusters aleatoriamente

clusterizar.fit(vetor)  #coloco algoritmo para trabalhar/analisar os dados -
#achar os padrões

nome_clusterizacoes = clusterizar.labels_  #cria variável para salvar as
#clusterizacoes

base_dados["Clusters"] = nome_clusterizacoes  #cria nova coluna no arquivo com
#os valores da clusterizacao

""" Graficos de Dispersão """
plt.figure( figsize=(18, 8))
plt.scatter(base_dados["TC"],base_dados["Clusters"], color='red', label="TC")
plt.scatter(base_dados["CT"],base_dados["Clusters"], color='blue',
label="CT") plt.scatter(base_dados["CU"],base_dados["Clusters"],
color='purple', label="CU")
plt.scatter(base_dados["LT"],base_dados["Clusters"], color='orange',
label="LT") plt.title("Colunas x Clusters") plt.legend() plt.show()

""" Numero de elementos em cada cluster """
base_dados["Clusters"].plot(kind="hist", title="Elementos de cada cluster")
```





Considerações:

Primeiramente foi feita uma análise superficial dos dados a partir da variável “estatística” e a partir disso foi possível ver que os dados não estavam de extrema discrepância, então não houve necessidade de normalizar os dados (as faixas dos dados estavam próximas).

O primeiro gráfico gerado coloca os clusters criados (eixo Y), em relação as colunas da base de dados. Isso foi feito para que fosse possível ver como está a distribuição em relação as classes. O segundo gráfico é apenas uma representação do número de amostras em cada cluster.

Para melhor compreensão, é possível ver que, em relação a coluna “TC” (que continha os dados *sport, EComm* e *holid*) a divisão criada pelos clusters está englobando duas classes em cada cluster. Exemplificando: a categoria “sport” está dentro do cluster 1.0 e 0; a categoria game está dentro do cluster 1.0 e 2.0;

Assim, conclui-se que a clusterização resultante foi aceitável, mas poderia ter sido melhor, havendo mais amostras.