Presentazione Progetto DT & ML

Appello di Febbraio 2018

Componenti del gruppo:
- Pietro Mattia Campi, 794157
- Marta Giltri, 795267

Introduzione generale

- **Obiettivo**: clusterizzare i paesi del mondo a seconda dell'inquinamento prodotto e della politica energetica.
- **Dataset utilizzati:**
  - *World Bank Data:* CO_2, energia rinnovabile, popolazione, polveri sottili PM2.5, prodotto interno lordo GDP.
  - *UNdata:* carbone, carburanti (benzina e diesel).
  - *Github:* elenco dei paesi con codice ISO 3166-1 Alpha-3.

Data Technologies

Dimensioni di Qualità - I

- **Prima dimensione:** Completeness.

Presentiamo qui, per comodità, solo la completezza di tabella per ogni dataset considerato indicata come percentuale, calcolata come:

$$\frac{\text{n° di valori not null}}{\text{n° di valori totali}}$$

Dataset	Percentuale Completezza
CO_2	76.87%
Carbone	92.58%
Carburanti	85.64%
Energia Rinnovabile	96.05%
Popolazione	94.68%
PIL	90.17%
PM2.5	48.92%
Elenco codici paesi	96.59%

Dimensioni di Qualità - II

- **Seconda dimensione:** Currency.
  - World Bank Data: aggiornati fino al 2016.
  - UNdata: aggiornati fino al 2015.
  - Github: aggiornato fino al 2017.

- **Terza dimensione:** Pertinency.

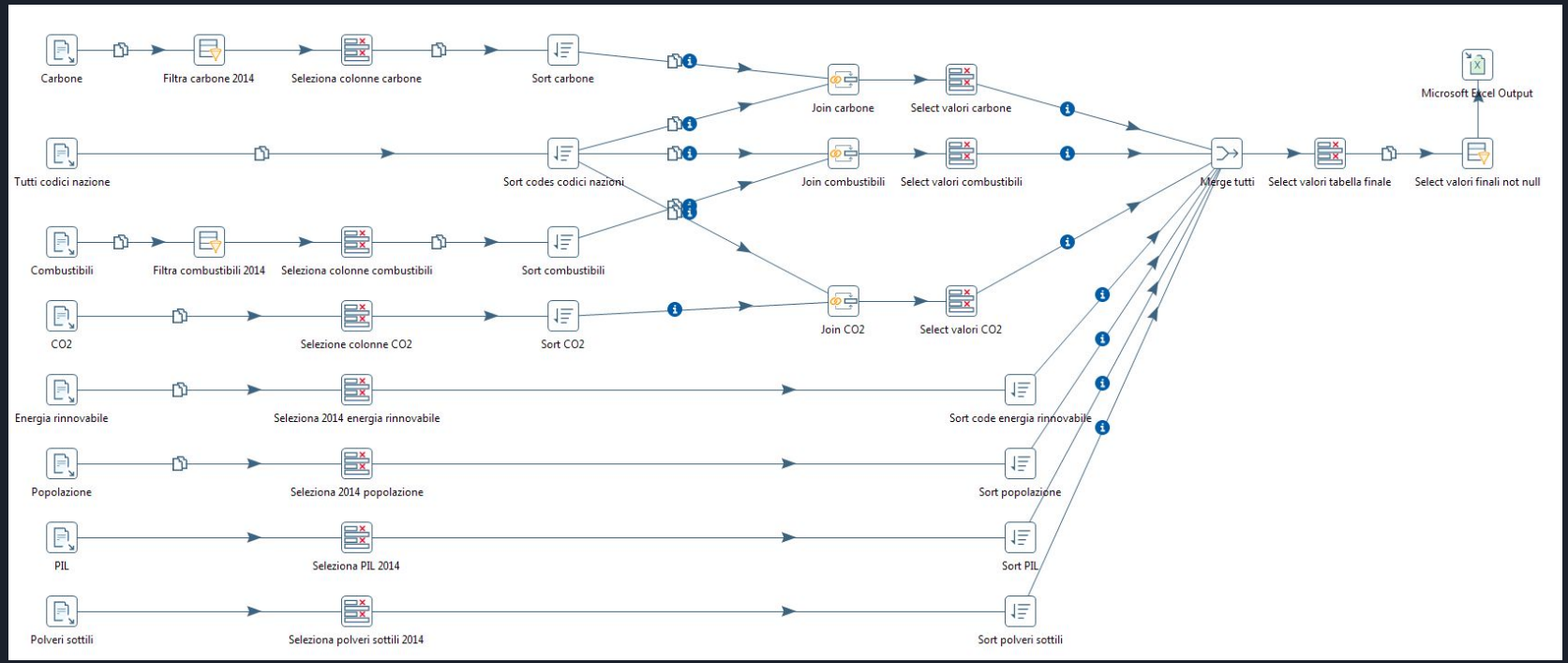
Calcolata tenendo conto delle colonne di dati utili.

Dataset	Colonne utili
CO_2	3 su 16
Carbone	4 su 6
Carburanti	4 su 6
Energia Rinnovabile	3 su 19
Popolazione	3 su 19
PIL	3 su 19
PM2.5	3 su 19
Elenco codici paesi	3 su 6

Data Integration - I

- Pulizia, Normalizzazione e Record Linkage su Elenco codici paesi e UNdata dataset (Python).
    - Rimozione delle stringhe “..” utilizzate per indicare i null.
    - Espansione sigle ed abbreviazioni presenti.
    - Eliminazione caratteri come parentesi e virgole.
    - Calcolo indice di similarità di Jaccard.
    - Associazione codice ISO3 sulla base di Jaccard.

Data Integration - II



Dimensioni di Qualità Dataset Finale

- **Prima dimensione:** Completeness.
  - Completezza di tabella: 100%.
  - Completezza di tupla: 153 su 153.
  - Completezza di attributi: 100% per tutti gli attributi.

- **Seconda dimensione:** Currency.
  - I dati risultano aggiornati per il 2014.

- **Terza dimensione:** Pertinency.
  - Colonne utili: 9 su 9.

Analisi Descrittiva Dataset Finale

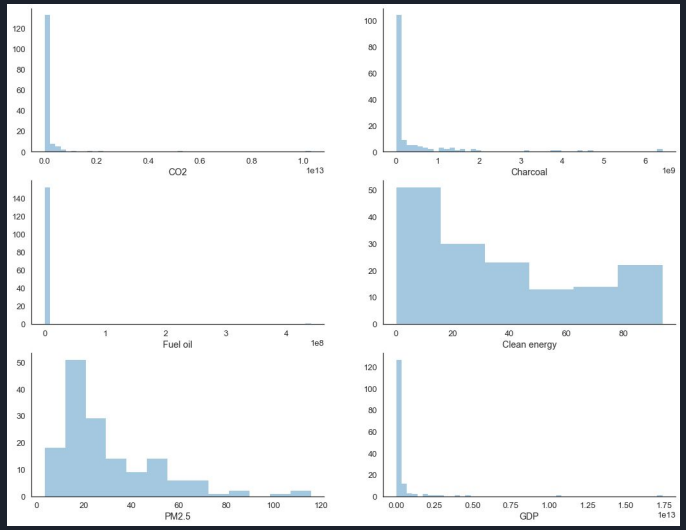
	Population	CO_2 prod. (kton)	Charcoal cons. (kton)	Fuel oil cons. (kton)	Renewable energy cons. (%)	PM2.5 (µg)	GDP (US $)
x	4.39e+07	1.991e+05	437.158	6735.986	35.687	30.244	4.32e+11
σ	1.56e+08	9.599e+05	1056.316	22262.31	29.877	21.075	1.74e+12
min	7.28e+04	6.234e+01	0.029	0.985	0.0	3.398	1.78e+09
max	1.36e+09	1.029e+07	6411.0	190050.0	93.859	115.872	1.73e+13

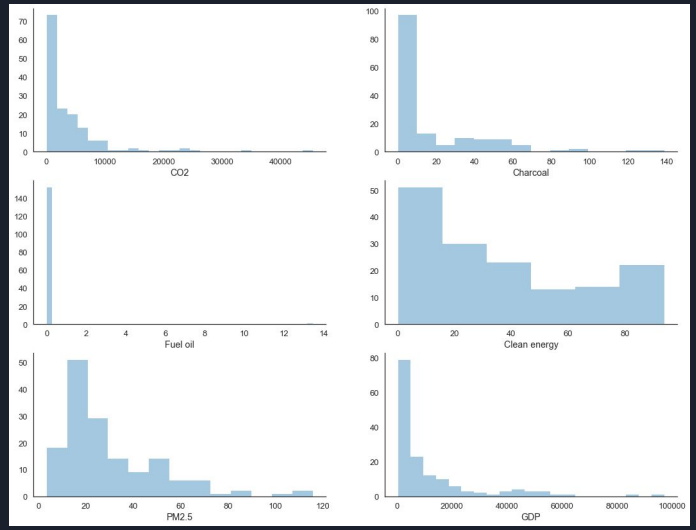
Machine Learning

Dataset Iniziale

Codice	Nome	Popolazione	CO_2	Charcoal	Fuel oil	Clean Energy	PM2.5	GDP
String	String	Int	Float	Float	Float	Float	Float	Int

Normalizzazione con la Popolazione

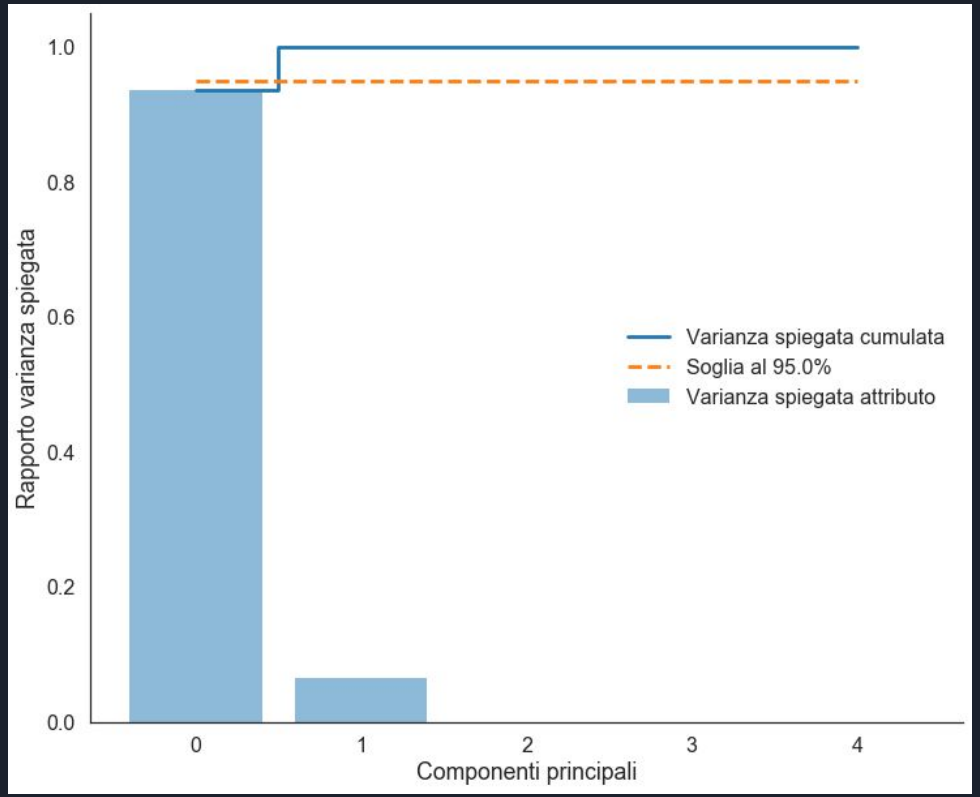




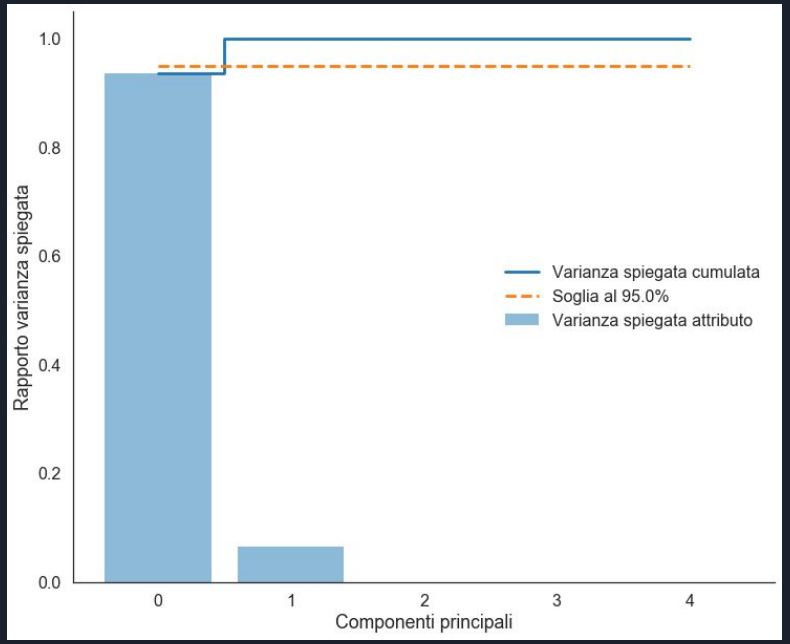
Statistiche Descrittive - Normalizzato

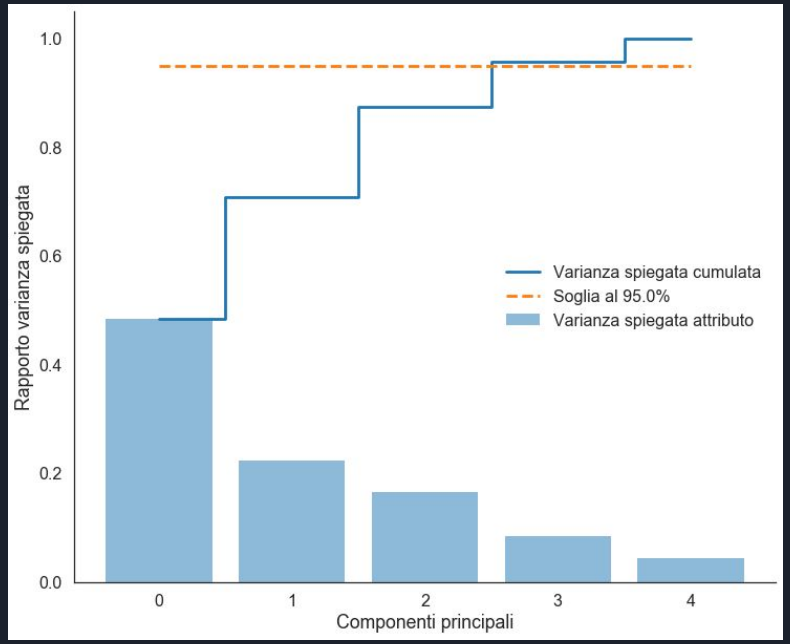
	CO_2 production	Charcoal	Fuel oil	Clean energy	PM2.5	GDP
x	4190.1	17.17	0.088	35.7	30.24	11757.0
σ	6372.36	25.6	1.09	29.9	21.1	16910.0
min	44.48	0.009	5.18e-07	0.0	3.40	312.75
max	45423.24	138.78	13.45	93.86	115.87	97200.0

Analisi delle componenti principali

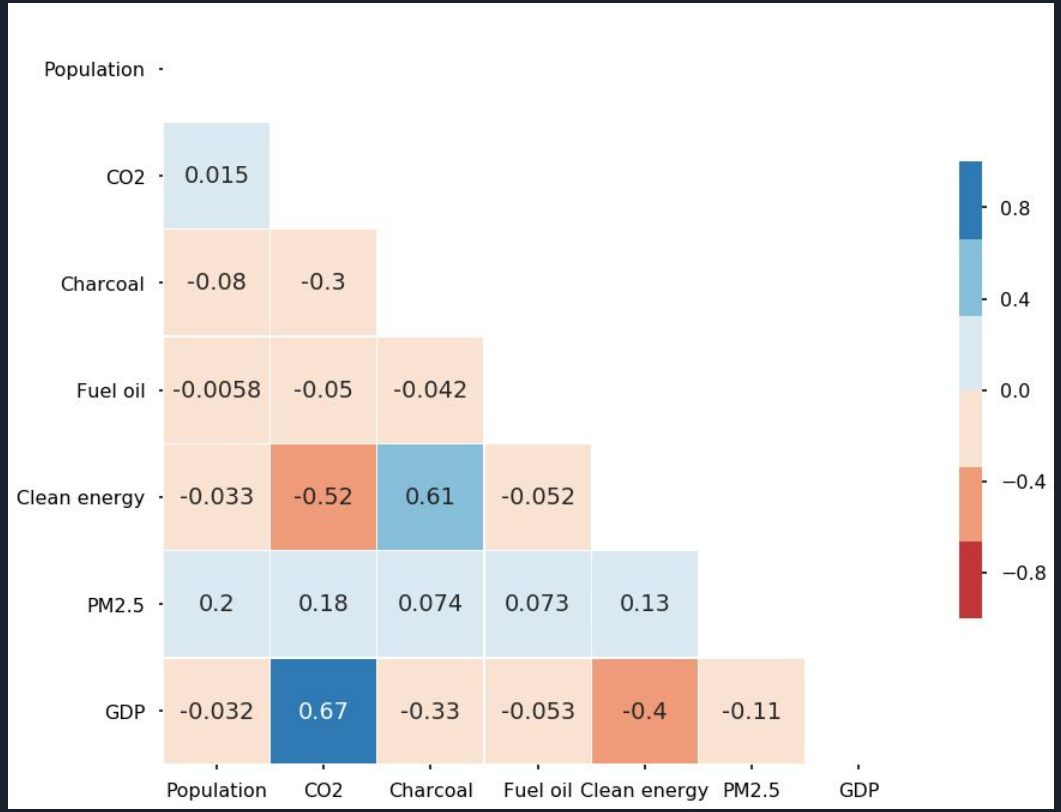


Standardizzazione



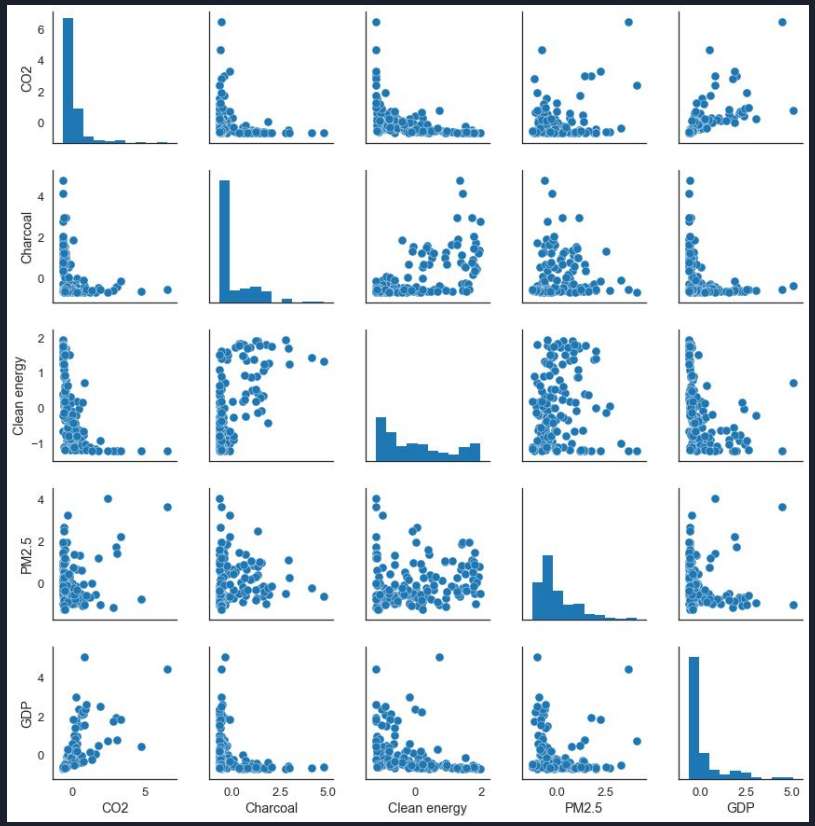


Analisi di correlazione attributi



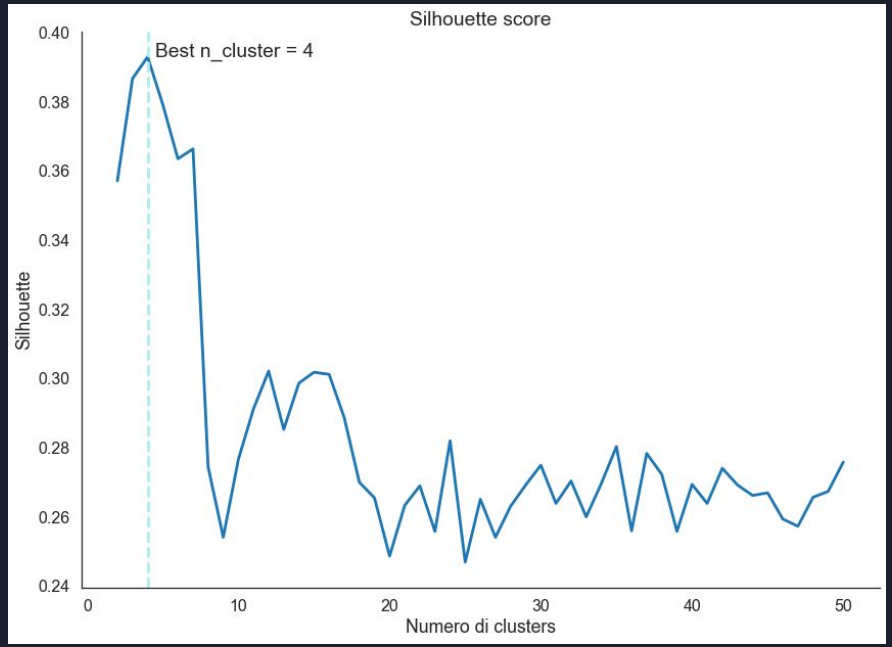
Distribuzioni accoppiate

Andamento a coppie degli attributi, utile per valutare la dispersione e la vicinanza in più di 3D



Modello K-Means

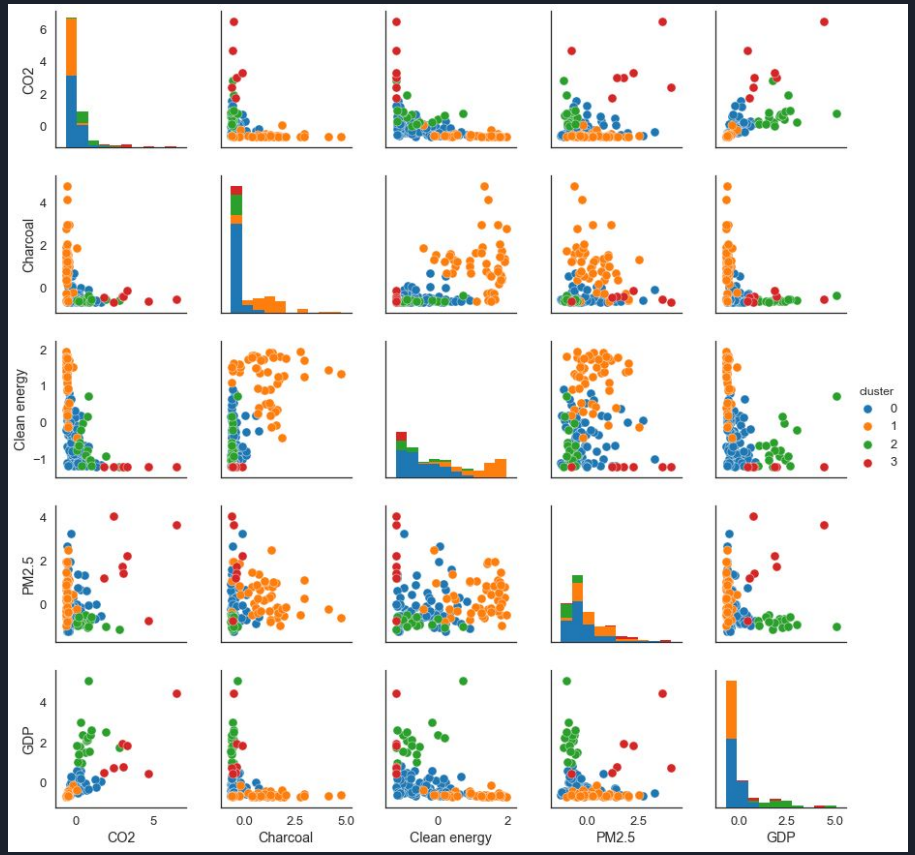
- Sul dataset normalizzato rispetto alla popolazione e standardizzato
- Uso di sklearn.clustering.KMeans
  - Da 2 a 50 cluster
  - Selezione sulla base dell'indice di silhouette



4 clusters

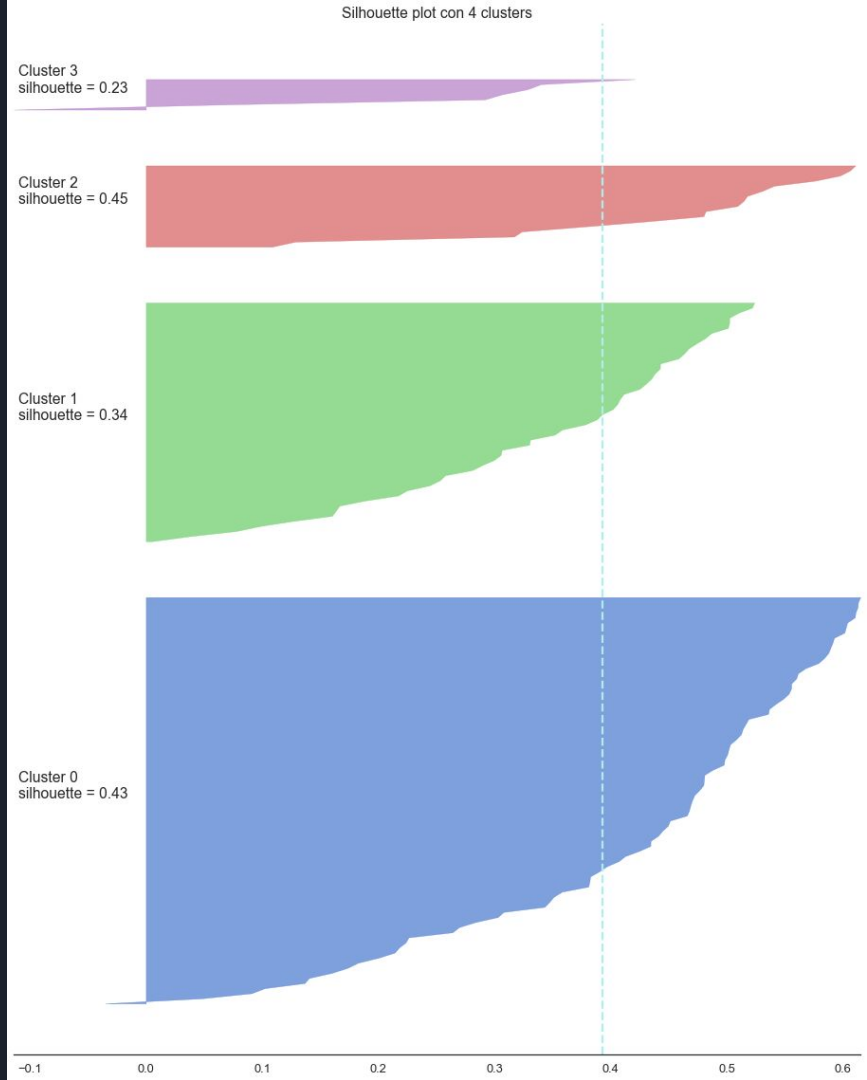
Silhouette media: 0.39

Visualizzazione dei cluster

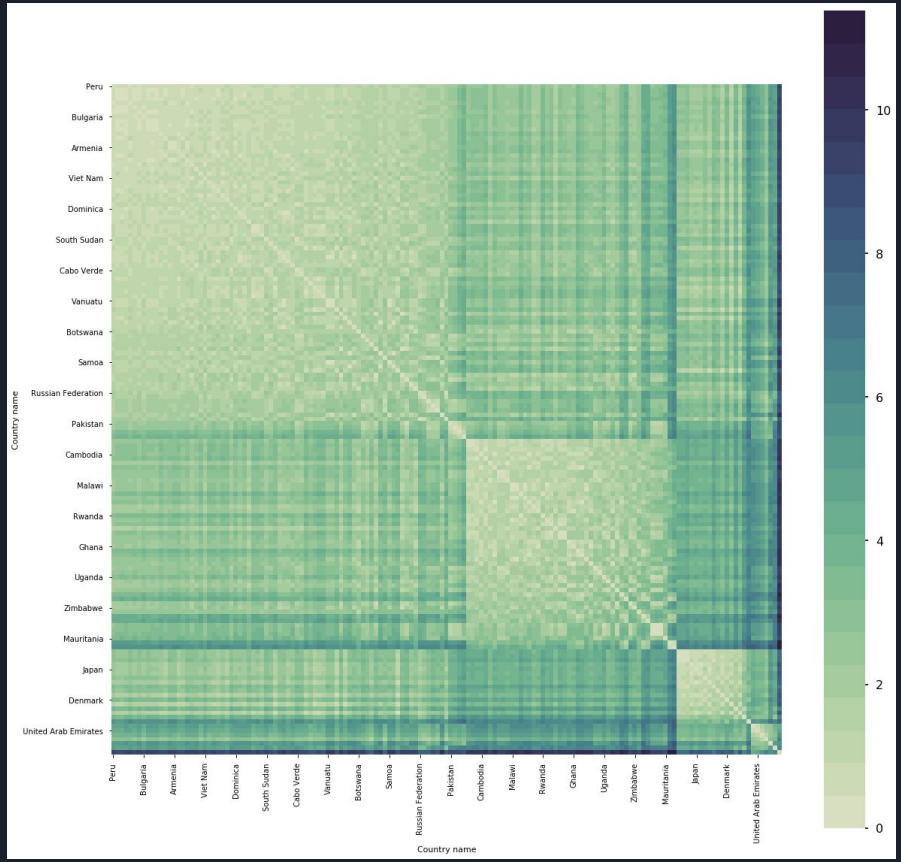


Silhouette per cluster

- Silhouette media 0.39, ma negativa per alcune istanze
- **Cluster 0**: paesi poveri medio-piccoli + Cina e India
- **Cluster 1**: paesi Africani e isole
- **Cluster 2**: il migliore, paesi ricchi dell'Eurozona + USA e Giappone
- **Cluster 3**: produttori di petrolio (medio-oriente)



Matrice di dissimilarità



Conclusioni

Conclusioni

Silhouette media **bassa** -> clusterizzazione non molto buona

Possibilità di aggiungere **altri attributi** più rappresentativi dei paesi

Analisi non completa: dati **mancanti** per circa **70 paesi**

Currency dei dati: analisi sui consumi del 2014, ma nel 2017?

Grazie dell’attenzione