EMISSIONS OF PESTICIDES IN THE E.U.

Katherine Atkins, Miryam Ochoa and Elvira Rodriguez

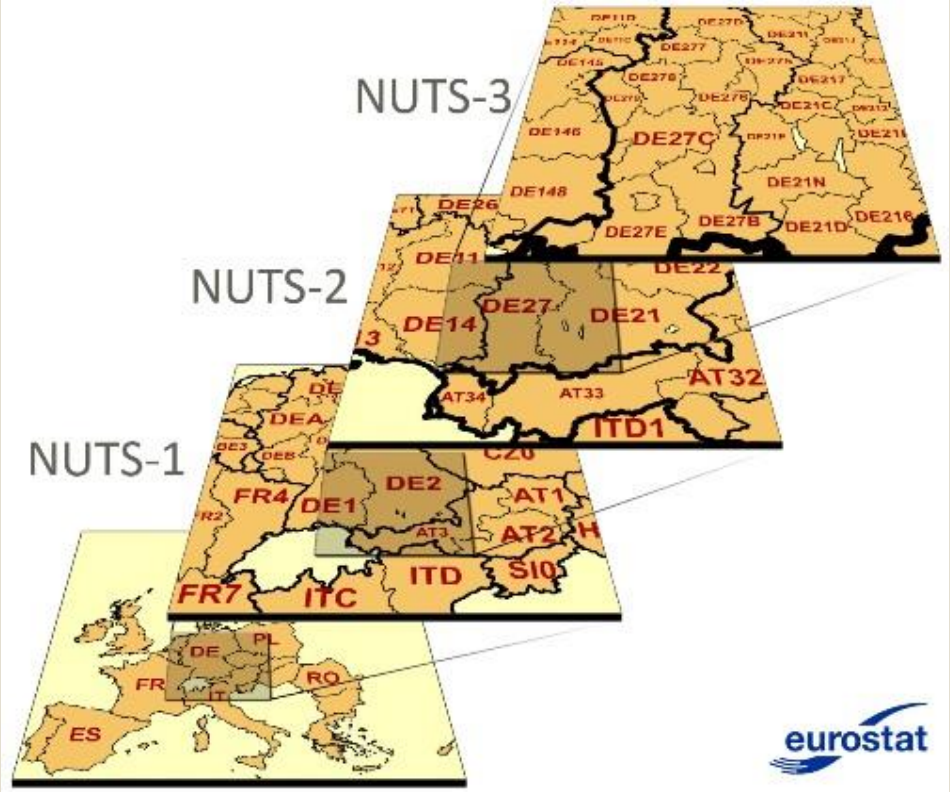
Project Objective:

“Emissions of pesticides in the European Union: a new regional-level dataset.” by Udias et al.

- Assessing the risks of pesticides is challenging because there are so many in use, and we don’t have enough detailed data on how and where they are being used.

- To address this problem, the study used for this project estimated pesticide use across the EU by analyzing detailed data on 150 different chemicals.

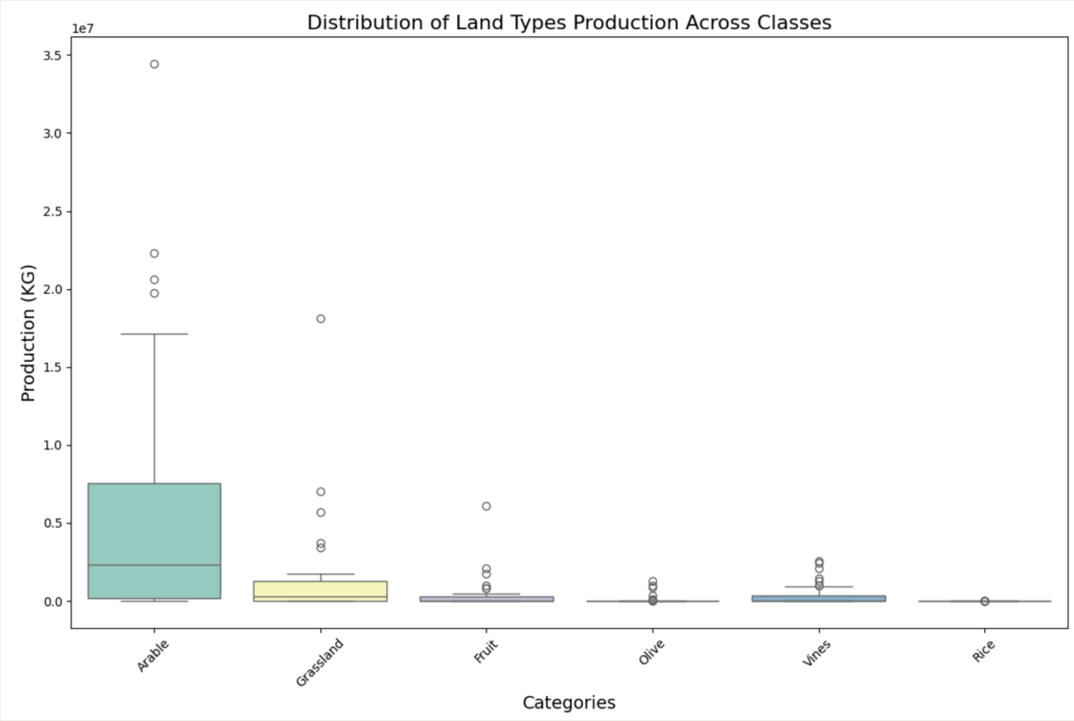
- The aim of this project is to use Unsupervised Machine Learning to find where significant use it.
  - Clustering techniques like K-Means and dimensionality reduction methods like PCA and t-SNE will aid in achieving our objective.



Data Understanding

```
Index(['COUNTRY', 'NUTS3', 'Categories_of_products',
       'Chemical_Class_Substance', 'ID_EUPDB', 'CAS', 'CIPAC',
       'Substances_common_names', 'Arab_KG', 'Fruit_KG', 'Oliv_KG', 'Vines_KG',
       'Grass_KG', 'Rice_KG', 'KG_TOT'],
      dtype='object')
```

- Country, NUTS3(level 3 units), category to which substance belongs, chemical class, EU id for pesticides, chemical abstract service council number identified, international number identifier, substance common name, land cover emissions.




```
[2]: data = pd.read_csv('pesticideActiveSubtances.csv')
[3]: data
```

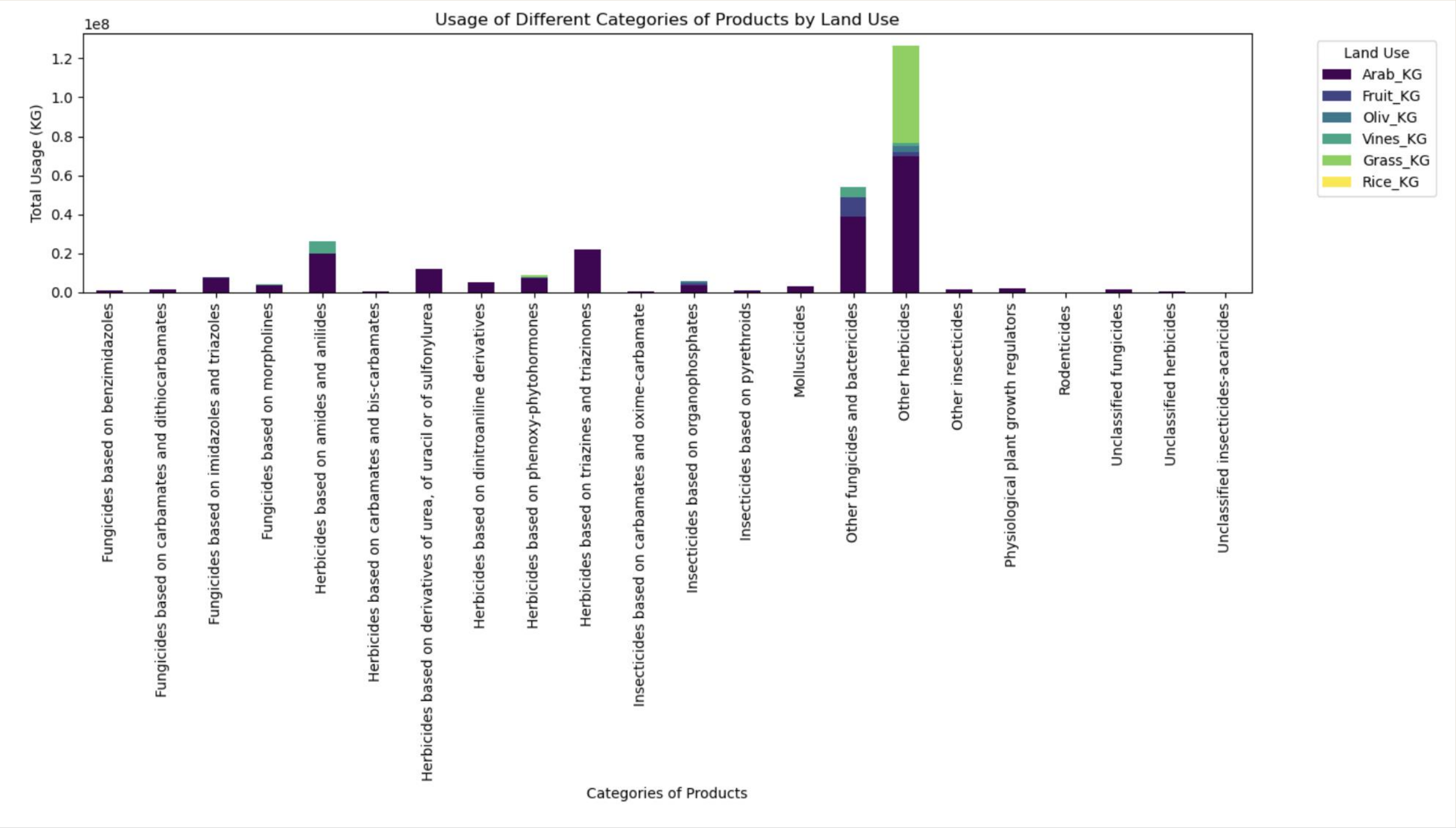
[3]:

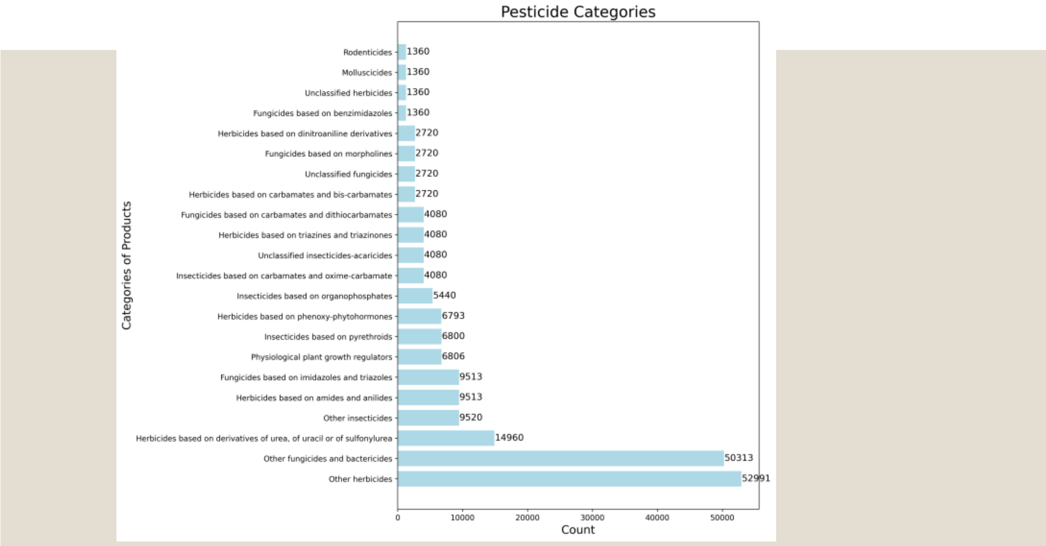
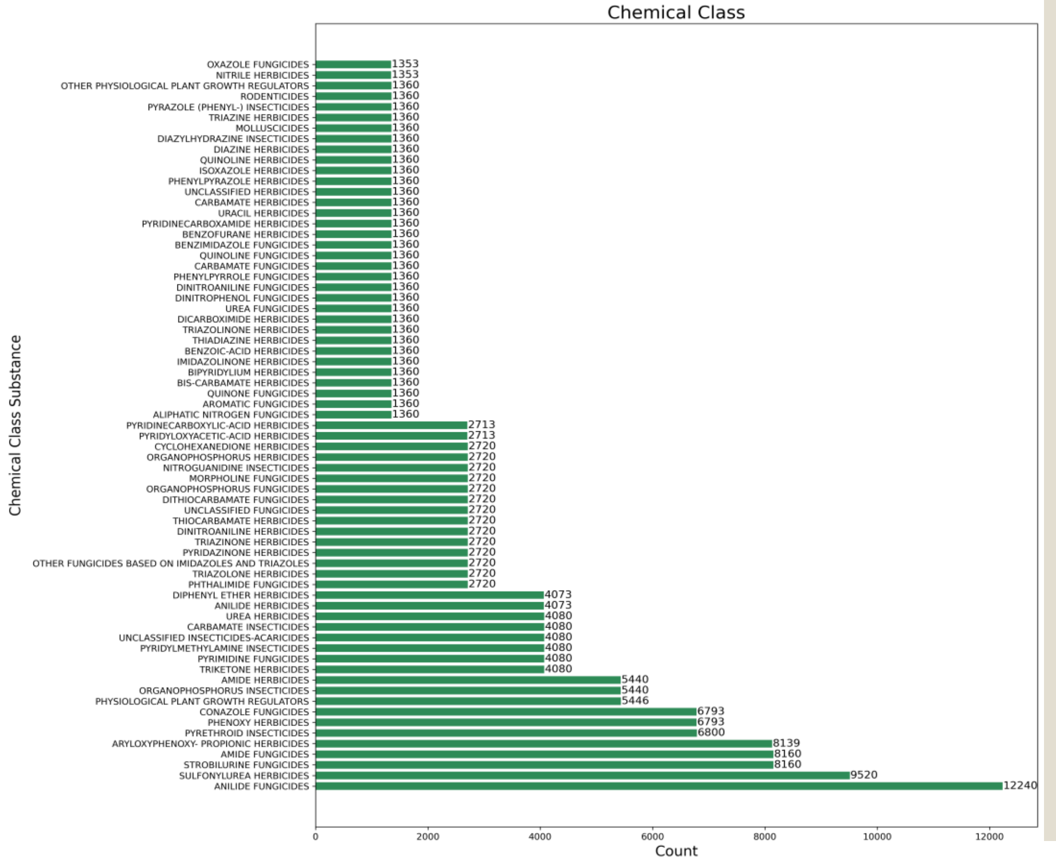
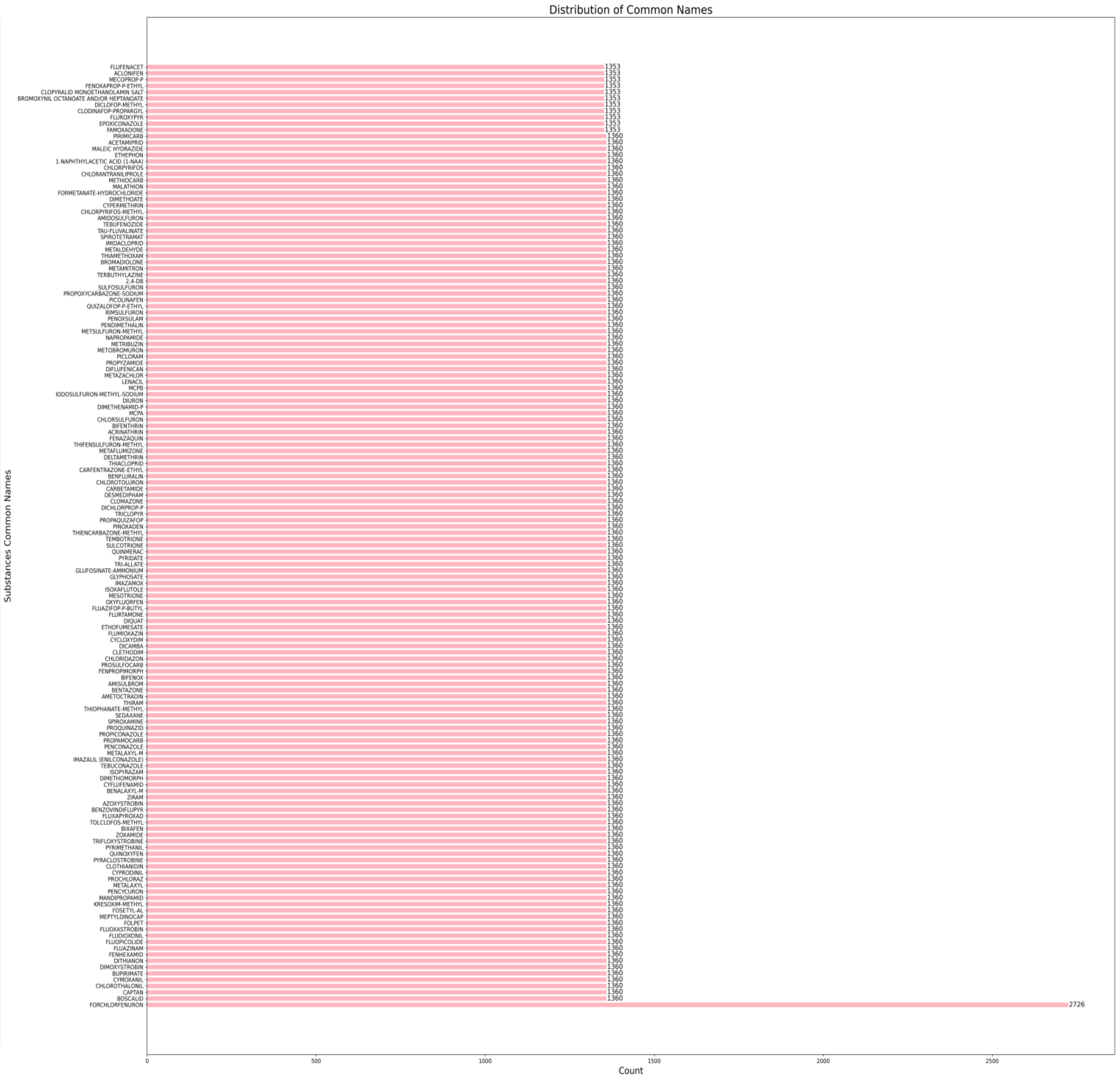
	COUNTRY	NUTS3	Categories_of_products	Chemical_Class_Substance	ID_EUPDB	CAS	CIPAC	Substances_common_names	Arab_KG	Fruit_KG	Oliv_KG
0	AT	AT111	Other fungicides and bactericides	ANILIDE FUNGICIDES	1040	188425-85-6	673.0	BOSCALID	760.385094	0.000000	0.0
1	AT	AT112	Other fungicides and bactericides	ANILIDE FUNGICIDES	1040	188425-85-6	673.0	BOSCALID	2702.234726	0.000000	0.0
2	AT	AT113	Other fungicides and bactericides	ANILIDE FUNGICIDES	1040	188425-85-6	673.0	BOSCALID	1521.092730	0.000000	0.0
3	AT	AT121	Other fungicides and bactericides	ANILIDE FUNGICIDES	1040	188425-85-6	673.0	BOSCALID	2262.038900	0.000000	0.0
4	AT	AT122	Other fungicides and bactericides	ANILIDE FUNGICIDES	1040	188425-85-6	673.0	BOSCALID	702.788789	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...
205284	UK	UKN01	Physiological plant growth regulators	OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS	856	86-87-3	313.0	1-NAPHTHYLACETIC ACID (1-NAA)	0.000000	0.000000	0.0
205285	UK	UKN02	Physiological plant growth regulators	OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS	856	86-87-3	313.0	1-NAPHTHYLACETIC ACID (1-NAA)	0.001873	0.000000	0.0
205286	UK	UKN03	Physiological plant growth regulators	OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS	856	86-87-3	313.0	1-NAPHTHYLACETIC ACID (1-NAA)	0.004087	0.397035	0.0
205287	UK	UKN04	Physiological plant growth regulators	OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS	856	86-87-3	313.0	1-NAPHTHYLACETIC ACID (1-NAA)	0.003429	0.000000	0.0
205288	UK	UKN05	Physiological plant growth regulators	OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS	856	86-87-3	313.0	1-NAPHTHYLACETIC ACID (1-NAA)	0.000909	0.469333	0.0

205289 rows × 15 columns

Data Understanding

- 35 Countries
- Six land cover classes
  - Arable Land
  - Fruit Trees
  - Grassland
  - Olive Groves
  - Vineyards
  - Rice Fields
- 22 Categories of Products




Pesticide Categories
Categories of Products
Rodenticides 1360
Molluscicides 1360
Unclassified herbicides 1360
Fungicides based on benzimidazoles 1360
Herbicides based on dinitroaniline derivatives 2720
Fungicides based on morpholines 2720
Unclassified fungicides 2720
Herbicides based on carbamates and bis-carbamates 2720
Fungicides based on carbamates and dithiocarbamates 4080
Herbicides based on triazines and triazinones 4080
Unclassified insecticides-acaricides 4080
Insecticides based on carbamates and oxime-carbamate 4080
Insecticides based on organophosphates 5440
Herbicides based on phenoxy-phytohormones 6793
Insecticides based on pyrethroids 6800
Physiological plant growth regulators 6806
Fungicides based on imidazoles and triazoles 9513
Herbicides based on amides and anilides 9513
Other insecticides 9520
Herbicides based on derivatives of urea, of uracil or of sulfonylurea 14960
Other fungicides and bactericides 50313
Other herbicides 52991
0
10000
20000
30000
40000
50000
Count

Chemical Class
Chemical Class Substance
OXAZOLE FUNGICIDES 1353
NITRILE HERBICIDES 1353
OTHER PHYSIOLOGICAL PLANT GROWTH REGULATORS 1360
RODENTICIDES 1360
PYRAZOLE (PHENYL-) INSECTICIDES 1360
TRIAZINE HERBICIDES 1360
MOLLUSCICIDES 1360
DIAZYLHYDRAZINE INSECTICIDES 1360
DIAZINE HERBICIDES 1360
QUINOLINE HERBICIDES 1360
ISOXAZOLE HERBICIDES 1360
PHENYLPYRAZOLE HERBICIDES 1360
UNCLASSIFIED HERBICIDES 1360
CARBAMATE HERBICIDES 1360
URACIL HERBICIDES 1360
PYRIDINECARBOXAMIDE HERBICIDES 1360
BENZOFURANE HERBICIDES 1360
BENZIMIDAZOLE FUNGICIDES 1360
QUINOLINE FUNGICIDES 1360
CARBAMATE FUNGICIDES 1360
PHENYLPYRROLE FUNGICIDES 1360
DINITROANILINE FUNGICIDES 1360
DINITROPHENOL FUNGICIDES 1360
UREA FUNGICIDES 1360
DICARBOXIMIDE FUNGICIDES 1360
TRIAZOLINONE HERBICIDES 1360
THIADIAZINE HERBICIDES 1360
BENZOIC-ACID HERBICIDES 1360
IMIDAZOLINONE HERBICIDES 1360
BIPYRIDYLIUM HERBICIDES 1360
BIS-CARBAMATE HERBICIDES 1360
QUINONE FUNGICIDES 1360
AROMATIC FUNGICIDES 1360
ALIPHATIC NITROGEN FUNGICIDES 1360
PYRIDINECARBOXYLIC-ACID HERBICIDES 2713
PYRIDYLOXYACETIC-ACID HERBICIDES 2713
CYCLOHEXANEDIONE HERBICIDES 2720
ORGANOPHOSPHORUS HERBICIDES 2720
NITROGUANIDINE INSECTICIDES 2720
MORPHOLINE FUNGICIDES 2720
ORGANOPHOSPHORUS FUNGICIDES 2720
DITHIOCARBAMATE FUNGICIDES 2720
UNCLASSIFIED FUNGICIDES 2720
THIOCARBAMATE HERBICIDES 2720
DINITROANILINE HERBICIDES 2720
TRIAZINONE HERBICIDES 2720
PYRIDAZINONE HERBICIDES 2720
OTHER FUNGICIDES BASED ON IMIDAZOLES AND TRIAZOLES 2720
TRIAZOLONE HERBICIDES 2720
PHTHALIMIDE FUNGICIDES 2720
DIPHENYL ETHER HERBICIDES 4073
ANILIDE HERBICIDES 4073
UREA HERBICIDES 4080
CARBAMATE INSECTICIDES 4080
UNCLASSIFIED INSECTICIDES-ACARICIDES 4080
PYRIDYLMETHYLAMINE INSECTICIDES 4080
PYRIMIDINE FUNGICIDES 4080
TRIKETONE HERBICIDES 4080
AMIDE HERBICIDES 5440
ORGANOPHOSPHORUS INSECTICIDES 5440
PHYSIOLOGICAL PLANT GROWTH REGULATORS 5446
CONAZOLE FUNGICIDES 6793
PHENOXY HERBICIDES 6793
PYRETHROID INSECTICIDES 6800
ARYLOXYPHENOXY- PROPIONIC HERBICIDES 8139
AMIDE FUNGICIDES 8160
STROBILURINE FUNGICIDES 8160
SULFONYLUREA HERBICIDES 9520
ANILIDE FUNGICIDES 12240
0
2000
4000
6000
8000
10000
12000
Count

Distribution of Common Names
Substances Common Names
FLUFENACET 1353
FORCHLORFENURON 2726
0
500
1000
1500
2000
2500
Count

Data Preprocessing

```
# Group by country and aggregate the emission columns for all six land covers
grouped_by_country = data.groupby('COUNTRY').aggregate({
    'Arab_KG': 'sum',
    'Fruit_KG': 'sum',
    'Oliv_KG': 'sum',
    'Vines_KG': 'sum',
    'Grass_KG': 'sum',
    'Rice_KG': 'sum',
    'KG_TOT': 'sum'
})
grouped_by_country = grouped_by_country.reset_index()
grouped_by_country
```

```
#remove unknown COUNTRY ISO code (e.g., 1-, 2-, etc)
grouped_by_country = grouped_by_country.iloc[6:]
grouped_by_country
```

- Grouped land use type by country
- Removed unknown country ISO codes
- Created a total_by_country feature to show usage of pesticide by country
- Converted from pandas to numpy array and found z-score of each country before conducting PCA

```
total_by_country = data.groupby(by="COUNTRY").aggregate({'Arab_KG':'sum','Fruit_KG':'sum','Oliv_KG':'sum',
                                                         'Vines_KG':'sum','Grass_KG':'sum','Rice_KG':'sum'})
total_by_country
```

```
#transform pandas df into numpy array
total_by_country_value = total_by_country.values
total_by_country_value
```

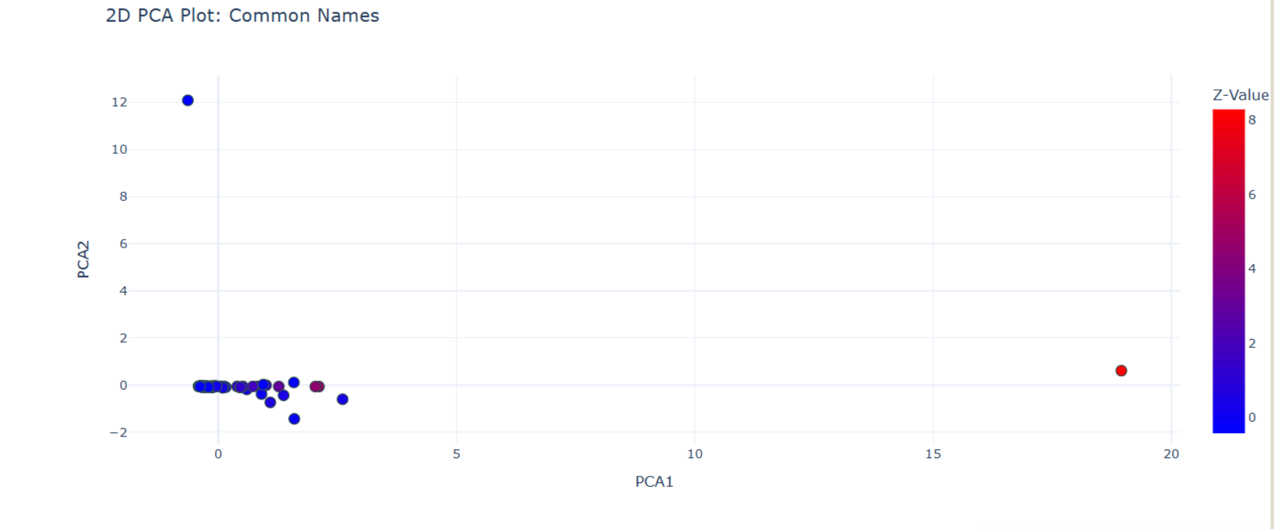
```
#import zscore package from scipy
from scipy import stats
#normalize the data using zscore
total_by_country_value_z = stats.zscore(total_by_country_value)
total_by_country_value_z
```

Data Exploration: Common Names

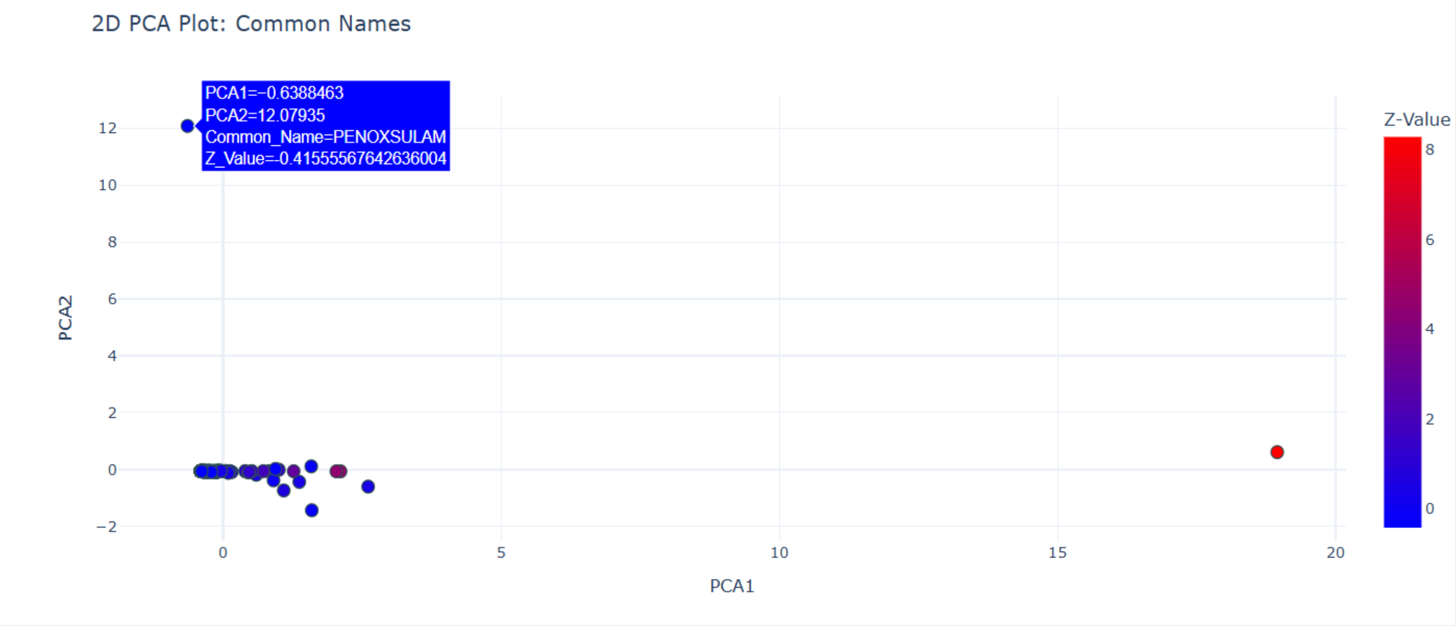
- (150, 6) shape.
- Z-score to normalize the data.
- Re-index
- Used PCA for dimension reduction

Substances_common_names	Arable	Grassland	Fruit	Olive	Vines	Rice
1-NAPHTHYLACETIC ACID (1-NAA)	7.158690e+01	0.000000	2645.632076	0.000000	0.000000	0.0
2,4-DB	3.104680e+05	0.000000	0.000000	0.000000	0.000000	0.0
ACETAMIPRID	5.965268e+04	0.000000	22690.306996	0.000000	3284.764587	0.0
ACLONIFEN	2.137123e+06	0.000000	0.000000	0.000000	0.000000	0.0
ACRINATHRIN	3.057213e+03	0.000000	0.000000	0.000000	4866.075996	0.0
...	...	...	...	...	...	...
TRI-ALLATE	2.354593e+06	0.000000	0.000000	0.000000	0.000000	0.0
TRICLOPYR	2.999254e+05	150149.531075	0.000000	0.000000	0.000000	0.0
TRIFLOXYSTROBINE	3.475378e+05	0.000000	52183.950174	6761.498656	56869.553173	0.0
ZIRAM	2.478025e+05	0.000000	321961.819484	0.000000	0.000000	0.0
ZOXAMIDE	6.000562e+04	0.000000	0.000000	0.000000	46345.448046	0.0

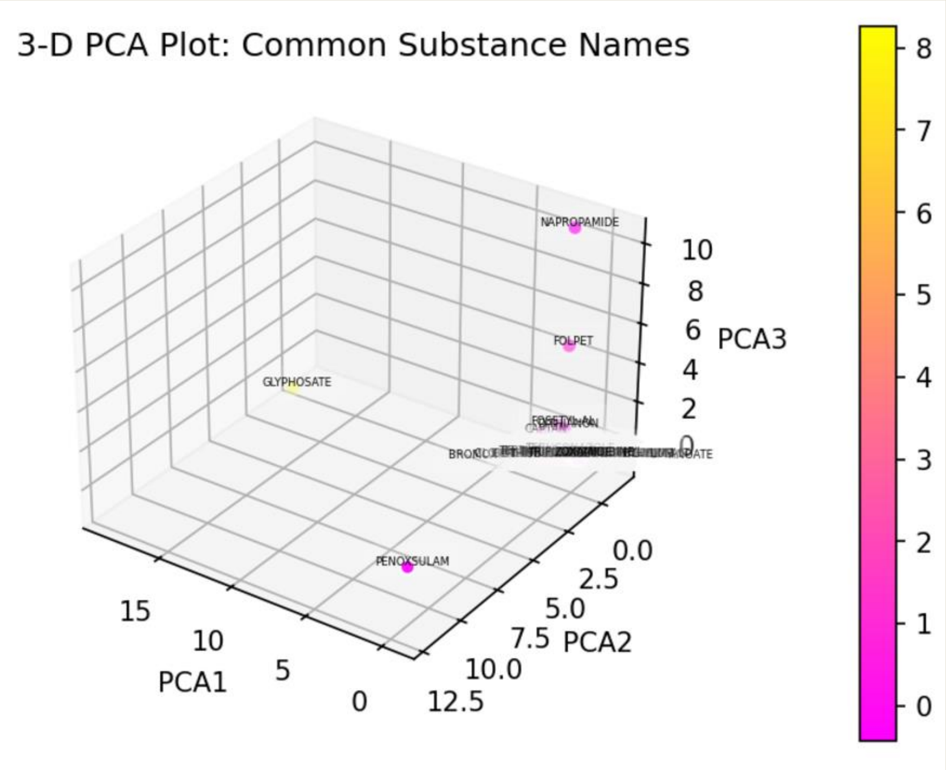
150 rows × 6 columns



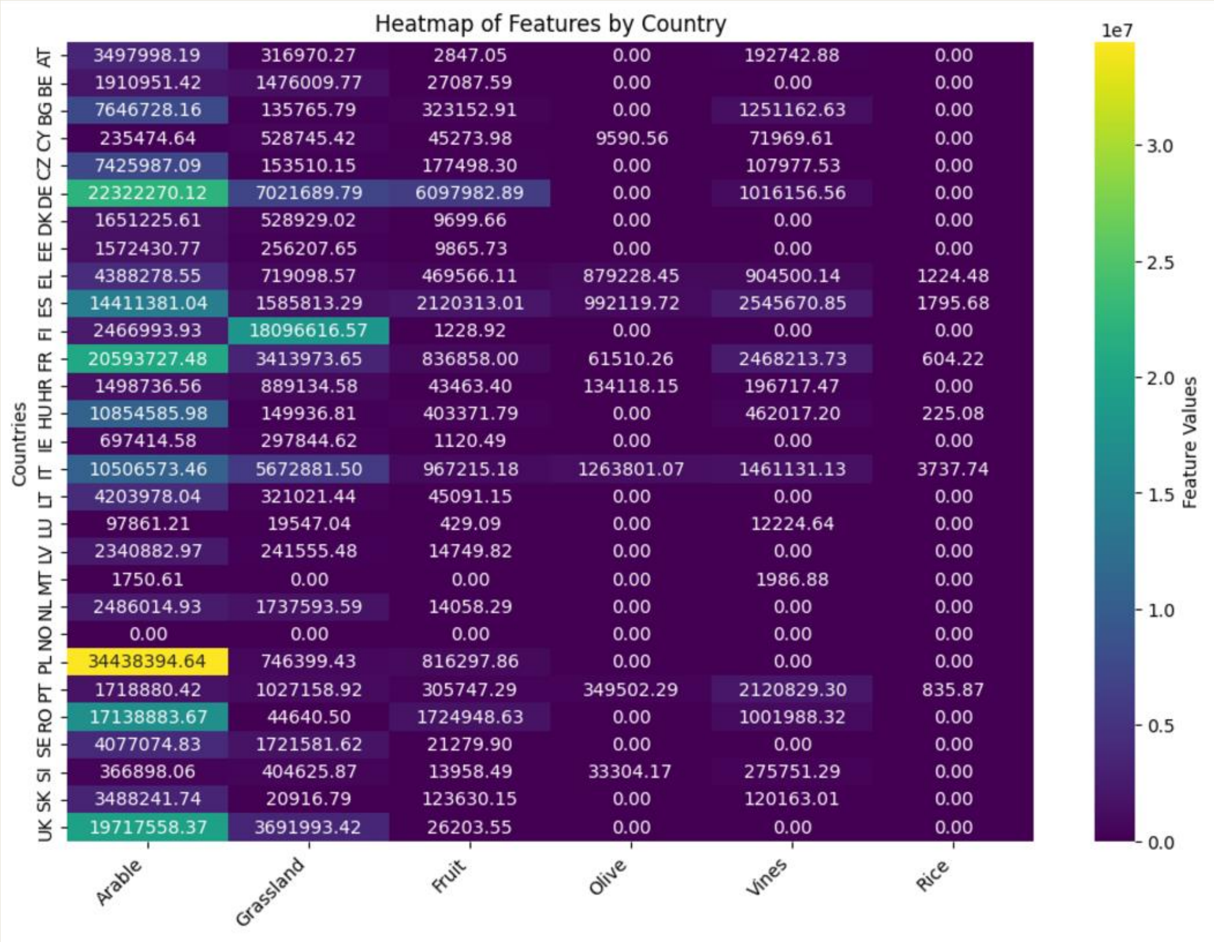

```
pca.explained_variance_ratio_
[83]:
array([0.44821953, 0.1669513 , 0.15771962])
[84]:
pca.explained_variance_ratio_.sum()
[84]:
np.float64(0.7728904443689487)
[85]:
1 - pca.explained_variance_ratio_.sum()
[85]:
np.float64(0.22710955563105129)
```

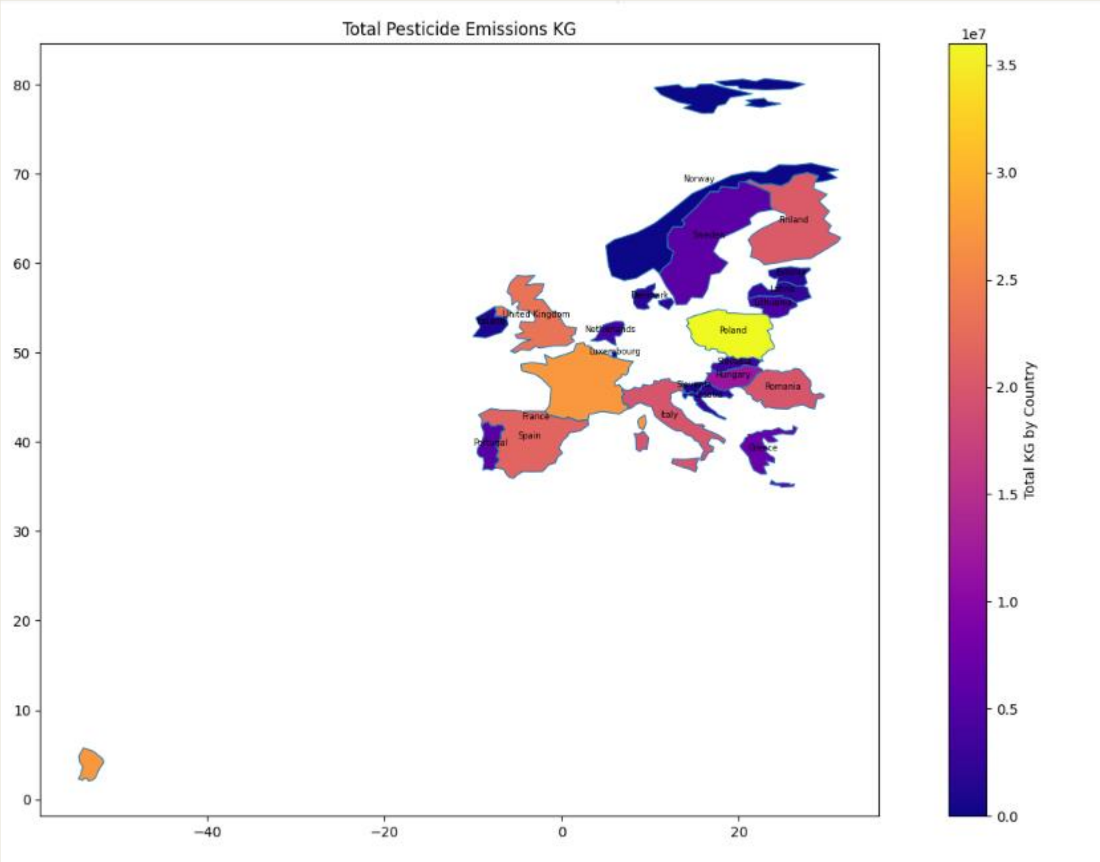

2D PCA Plot: Common Names
PCA1=−0.6388463
PCA2=12.07935
Common_Name=PENOXSULAM
Z_Value=-0.41555567642636004
Z-Value
PCA2
PCA1


2D PCA Plot: Common Names
PCA1=18.94883
PCA2=0.6062006
Common_Name=GLYPHOSATE
Z_Value=8.268777663572386
Z-Value
PCA2
PCA1

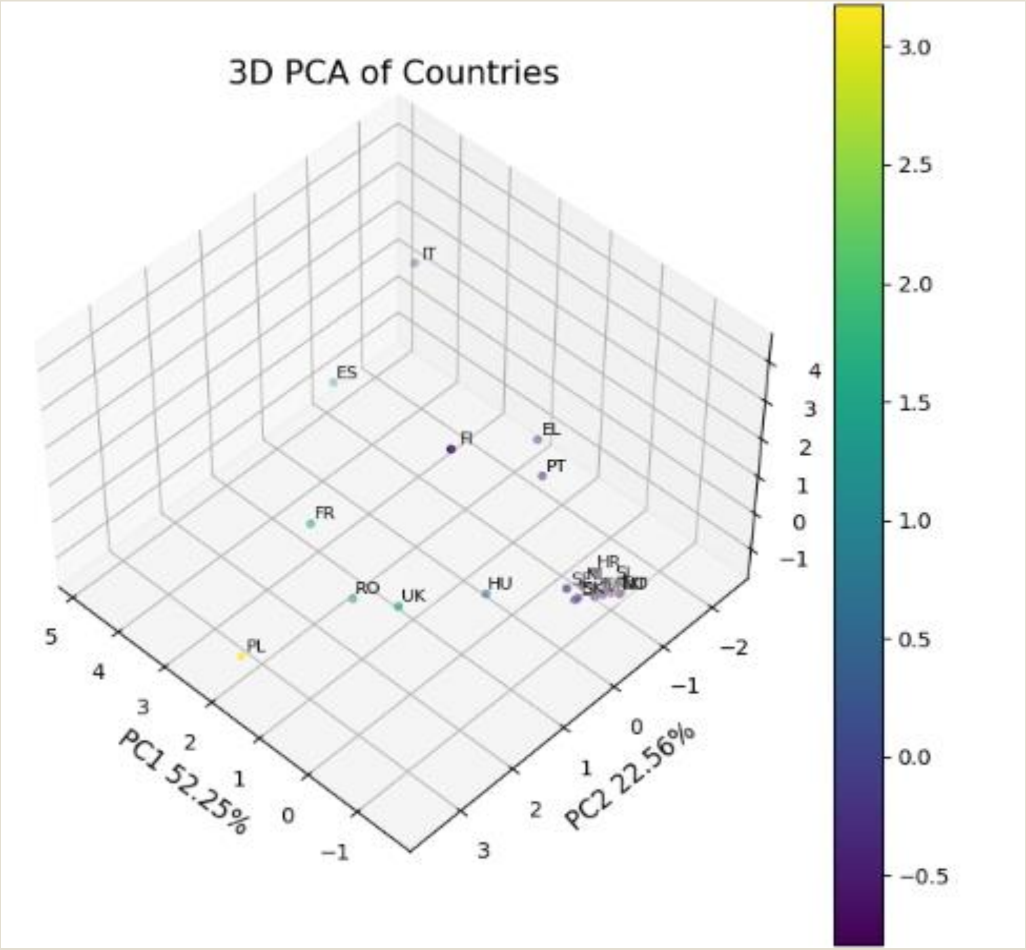

3-D PCA Plot: Common Substance Names
NAPROPAMIDE
FOLPET
GLYPHOSATE
PENOXSULAM
PCA1
PCA2
PCA3

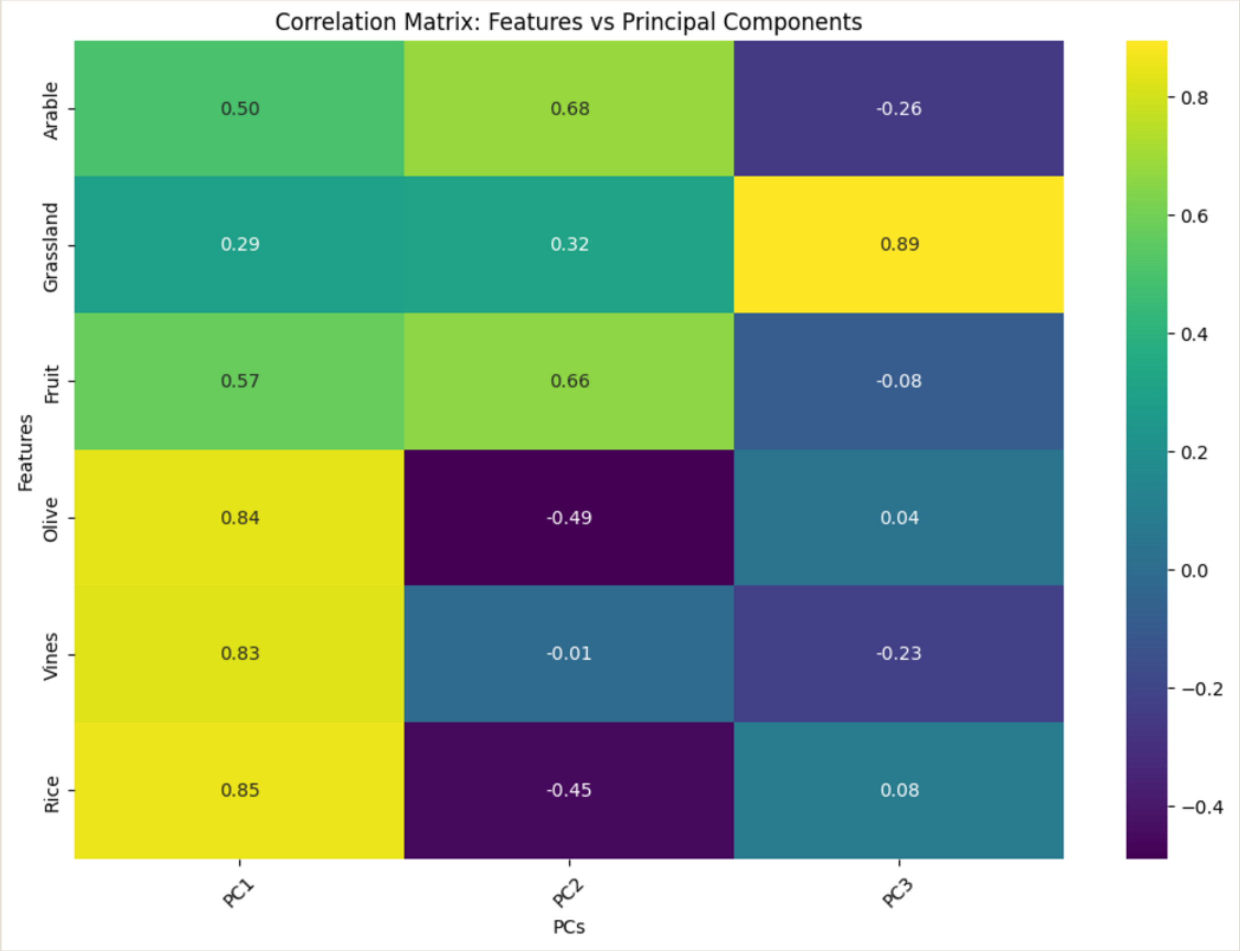
Data Processing





- Poland shows significant pesticide usage.
- Map was developed using GeoPandas and merging aggregated data onto the European shapefile. This shows the distribution of emission geospatially on the EU map.





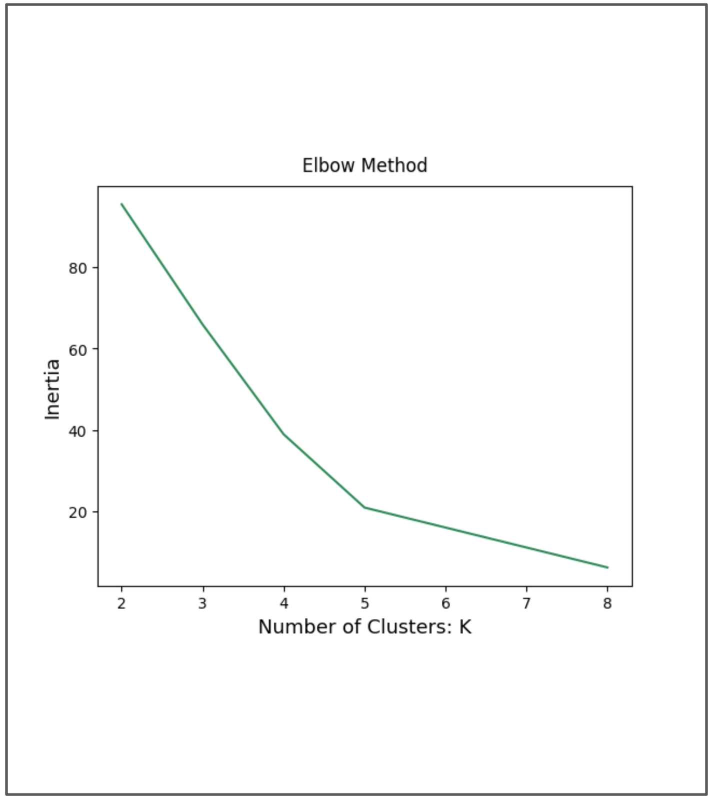
- The first three components explain 90.4% of the total variance in the dataset.
- Countries in this cluster may share similarities in agricultural practices, climate, or economic factors tied to the features
- Principal components focus on important features influencing emissions as explained by:

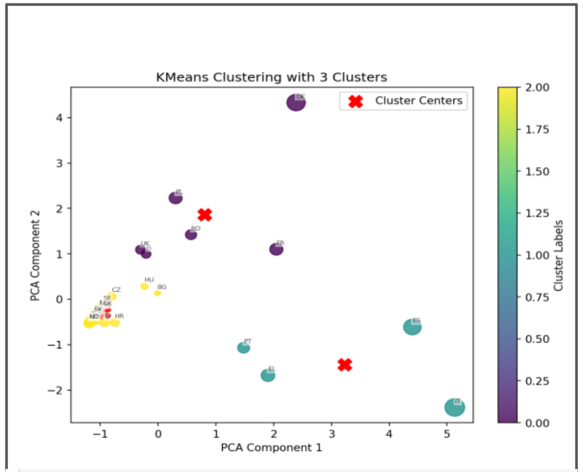
```
# variance explained in n component 3 pca?
pca_3.explained_variance_ratio_
```

```
array([0.52246419, 0.22556544, 0.15594409])
```

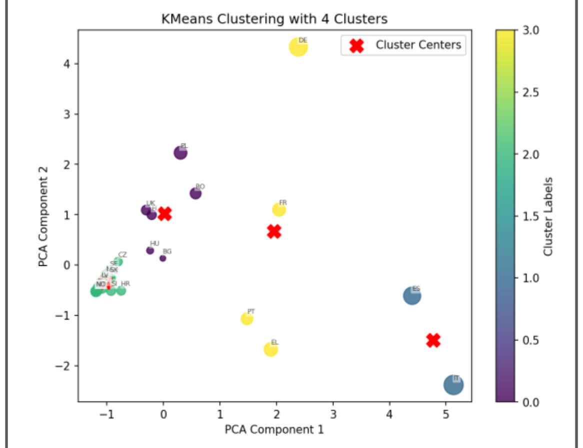
KMeans Clustering

- Why did we visualize k-means clustering results in PCA coordinates?

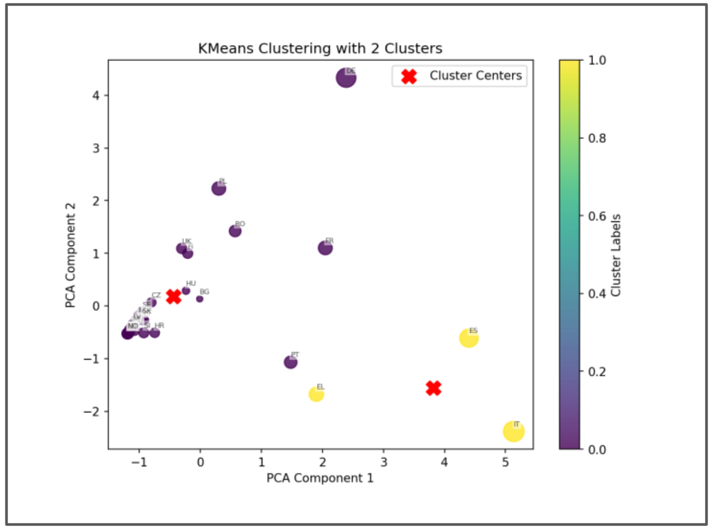


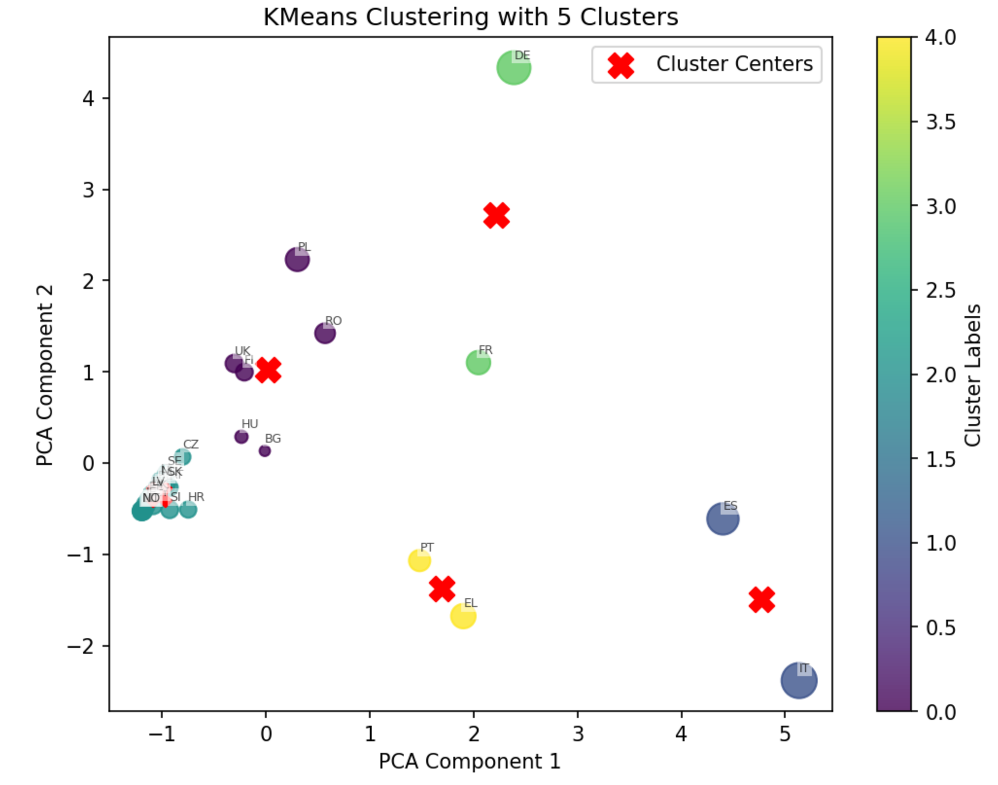


Silhouette Score for 3 clusters: 0.5850195644297902



Silhouette Score for 4 clusters: 0.5385298384276896

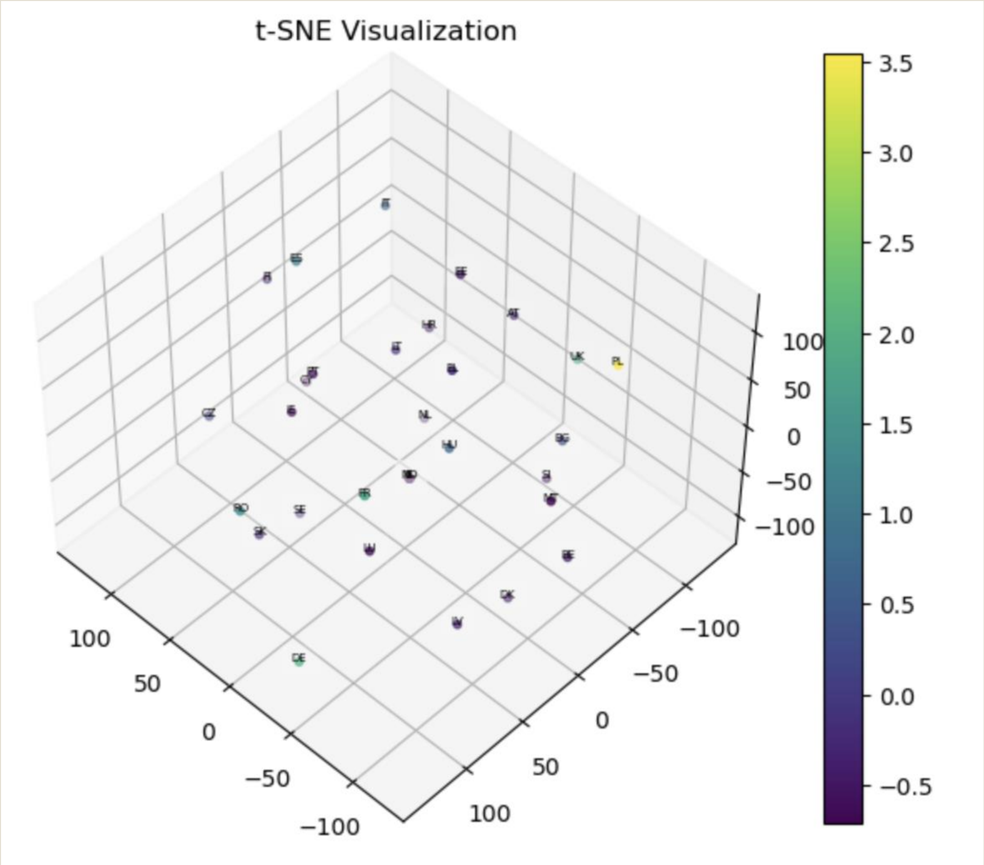




Silhouette Score for 5 clusters: 0.6560715839504054

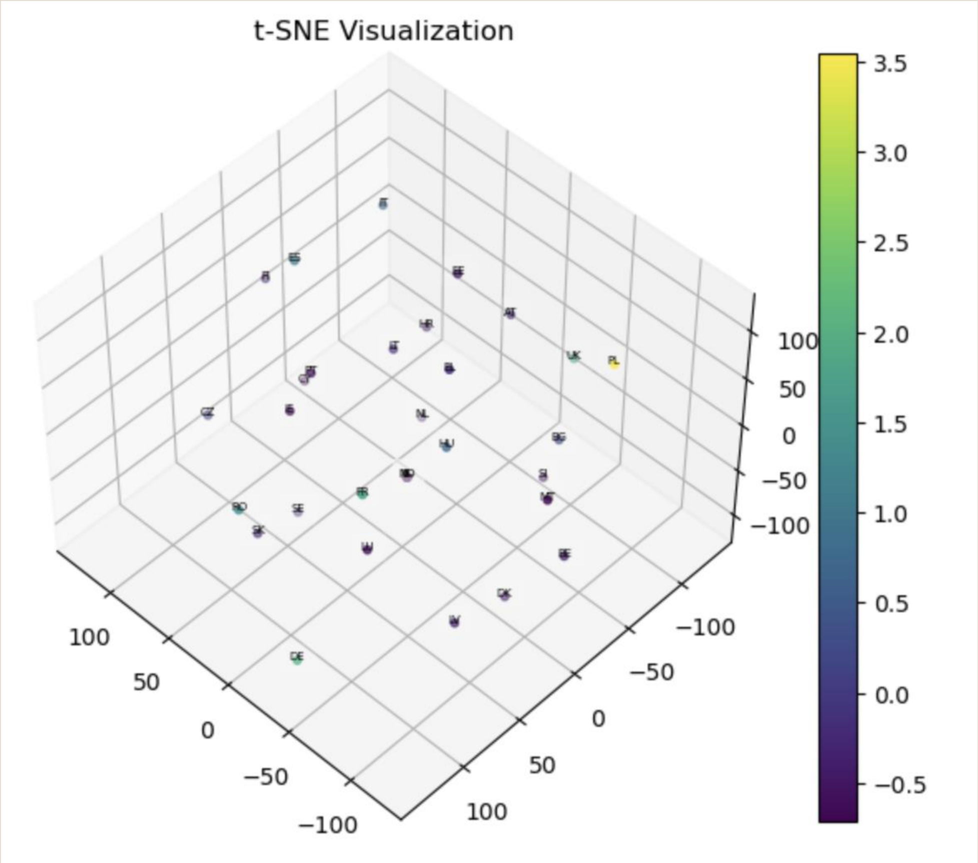
Clustering Improvement: Using TSNE

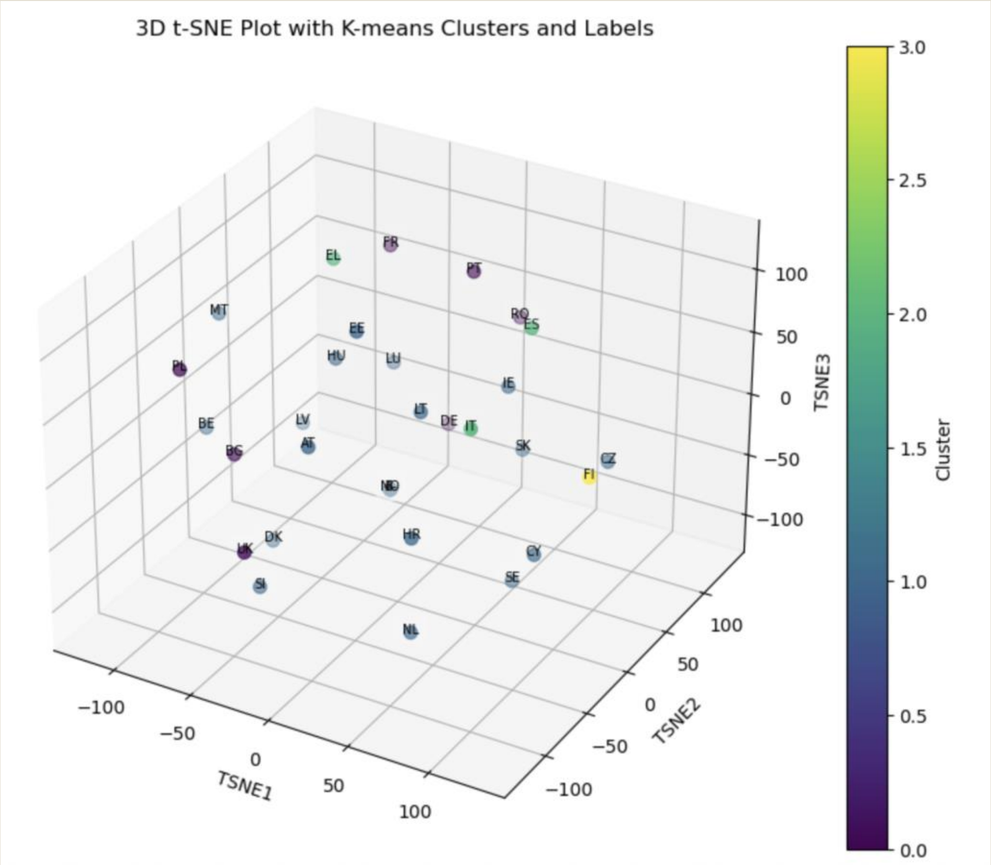
- *non-linear approach

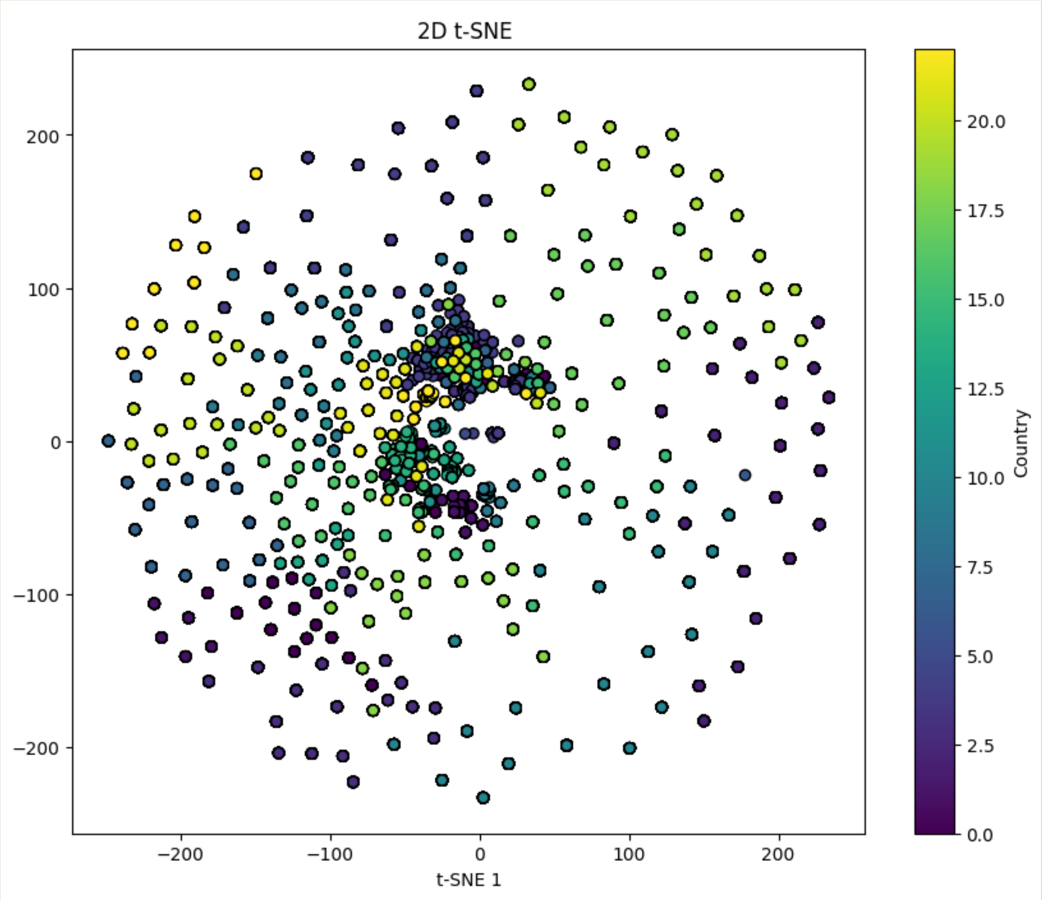


	TSNE1	TSNE2	TSNE3	Country
0	-6.373781	2.998195	-42.753185	1-
1	-6.373781	2.998195	-42.753185	2-
2	-6.373781	2.998195	-42.753185	3-
3	-6.373781	2.998195	-42.753185	4-
4	-6.373781	2.998195	-42.753185	5-
5	-6.373781	2.998195	-42.753185	6-
6	-6.092248	-89.616875	47.044476	AT
7	-120.396286	-8.934873	-26.851919	BE
8	-70.721169	-60.718315	0.374296	BG
9	93.459045	-13.063423	-48.040974	CY
10	99.712990	64.994820	-18.700502	CZ
11	-39.629089	135.375046	-80.076721	DE
12	-80.338486	-3.642405	-109.042831	DK
13	2.538164	-50.552940	116.710991	EE
14	-47.499779	12.551329	121.422134	EL
15	79.111206	7.836456	112.784737	ES
16	128.047958	-16.439491	28.601910	FI
17	-58.110081	97.621445	82.697525	FR
18	51.686378	-78.584175	-9.679397	HR
19	-47.582546	14.253402	41.510025	HU
20	42.623234	52.207787	28.846365	IE
21	87.413582	-78.310867	88.240509	IT
22	25.488886	-19.751251	44.481167	LT
23	-48.173866	83.259857	-1.200919	LU
24	-96.541092	60.485477	-55.420376	LV
25	-121.670044	10.099753	54.603489	MT
26	45.505329	-67.938820	-94.424751	NL
27	-6.373781	2.998195	-42.753185	NO
28	-89.352440	-85.829552	75.438400	PL
29	14.812978	62.283527	104.883064	PT
30	10.095901	129.819305	29.436085	RO
31	55.835361	34.134247	-114.128510	SE
32	-44.481133	-79.463760	-85.398384	SI
33	30.526255	94.891090	-52.077217	SK
34	-24.632862	-128.601639	-18.569016	UK

Plotting t-SNE to K-Means Clusters

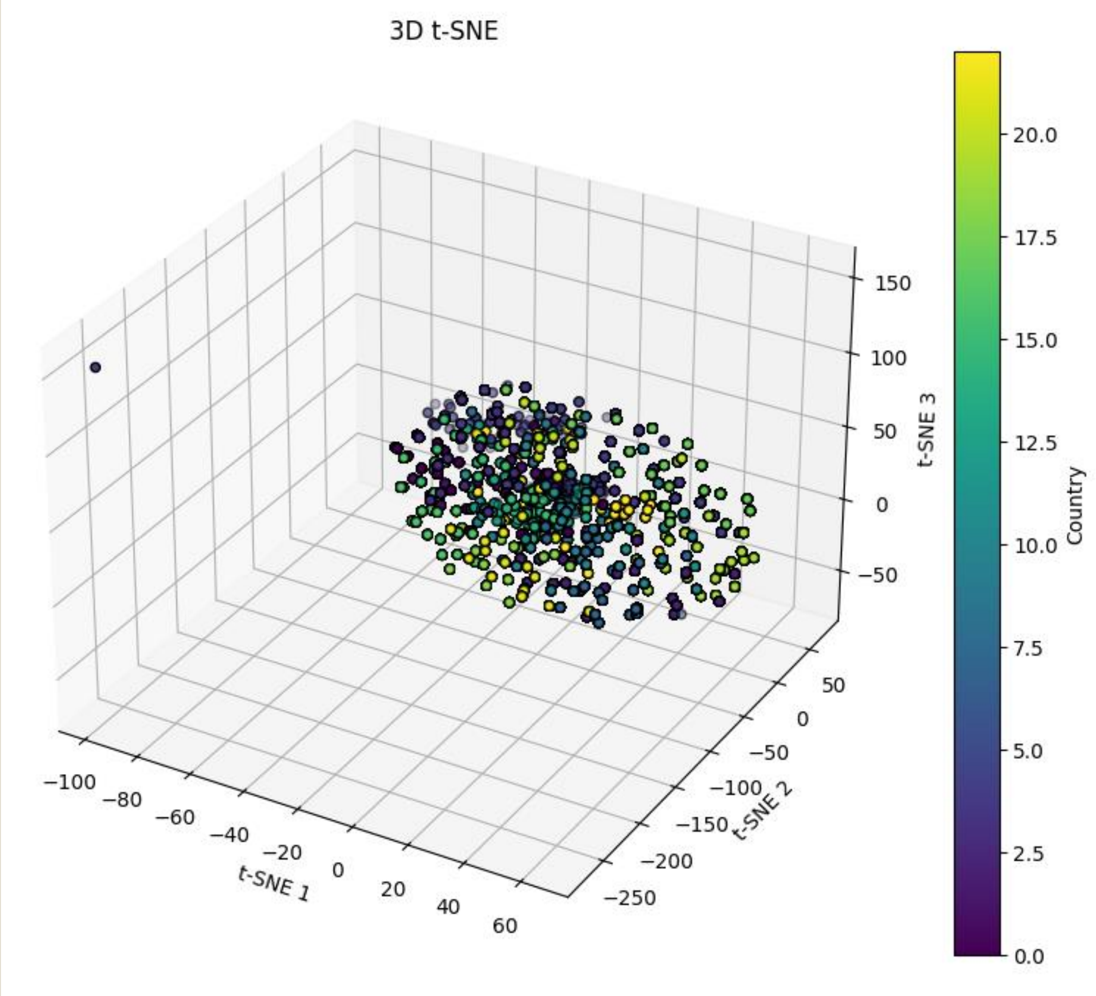







```
merged_df.shape

(92186, 9)
```



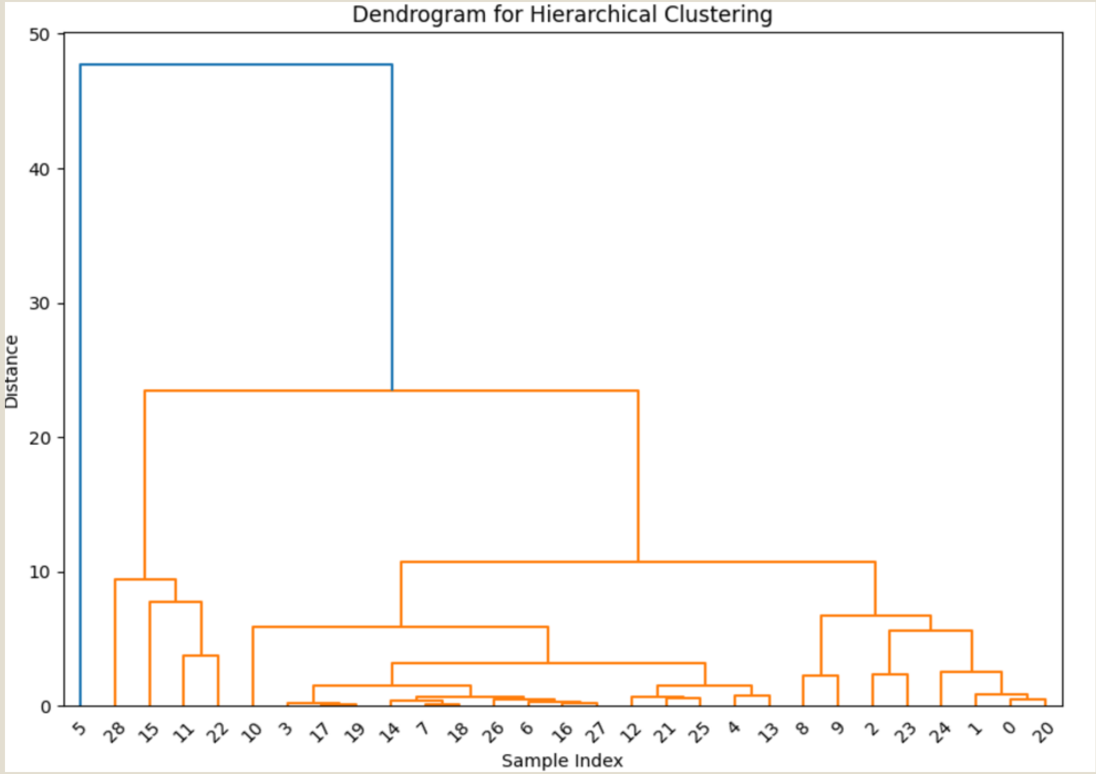
- Clustering display of a higher dimension.
- N_components
- Perplexity(It can be considered a measure of how many nearest neighbors are considered when computing the pairwise similarities of points.) We focused more local.

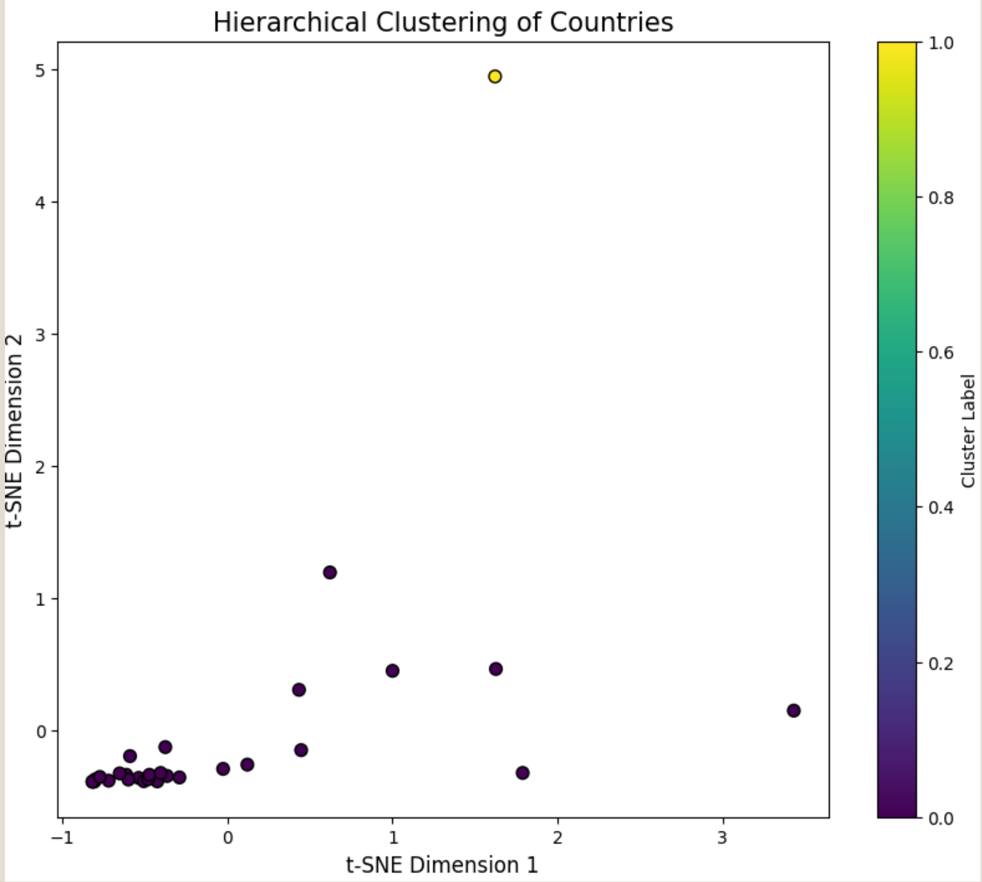
Hierarchical Clustering

- n_clusters
- metric='euclidean'
- Used for comparison to show the shape of the data

```
z_kg_by_country.shape
```

```
(29, 56)
```






```
Silhouette Score for 2 clusters: 0.8253041207093019
```

Conclusions/ Model Improvements

- Based on the Silhouette Scores of 0.825, the clustering used in this project is fairly good but could be improved using other clustering and tuning methods.
- While there is separation in the plotted points, there is ambiguity in the clustering as boundary is not clearly defined.
- A larger dataset or selecting different features and hypertuning would likely result in better clustering.
- Maps showing where and how much pesticides are likely used can be applied when studying their effects, like pollution in rivers.
- Understanding the application of pesticide usage and by country and land type can help suggest agricultural management, assess the ecotoxicology of arable land for future use, and develop sustainable policy.
- This helps fill the data gap and provide a clearer picture of pesticide use across Europe.

References

- KMeans — scikit-learn 1.5.2 documentation
- TSNE — scikit-learn 1.5.2 documentation
- Udias, A., Galimberti, F., Dorati, C., & Pistocchi, A. (2023, December 5). *Emissions of pesticides in the European Union: A new regional-level dataset*. Nature News. https://www.nature.com/articles/s41597-023-02753-4