

Breast Cancer Analysis

An analysis on the *Global Health Data Exchange*

DataSolve Group Members:

- Bernasconi Ariele
- Bollati Daniel
- Giacometti Carlo
- Gucciardo Valeria

Case Study

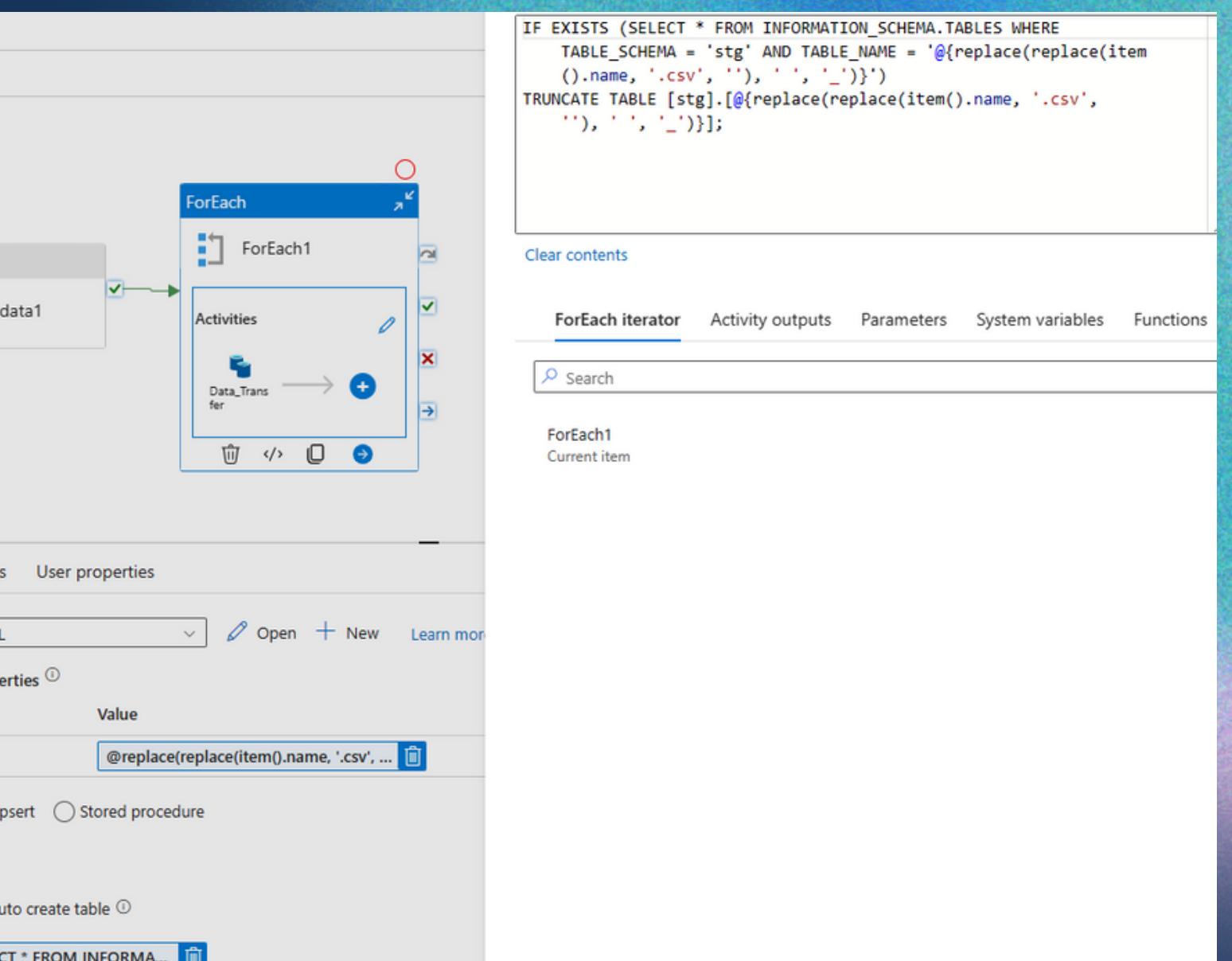
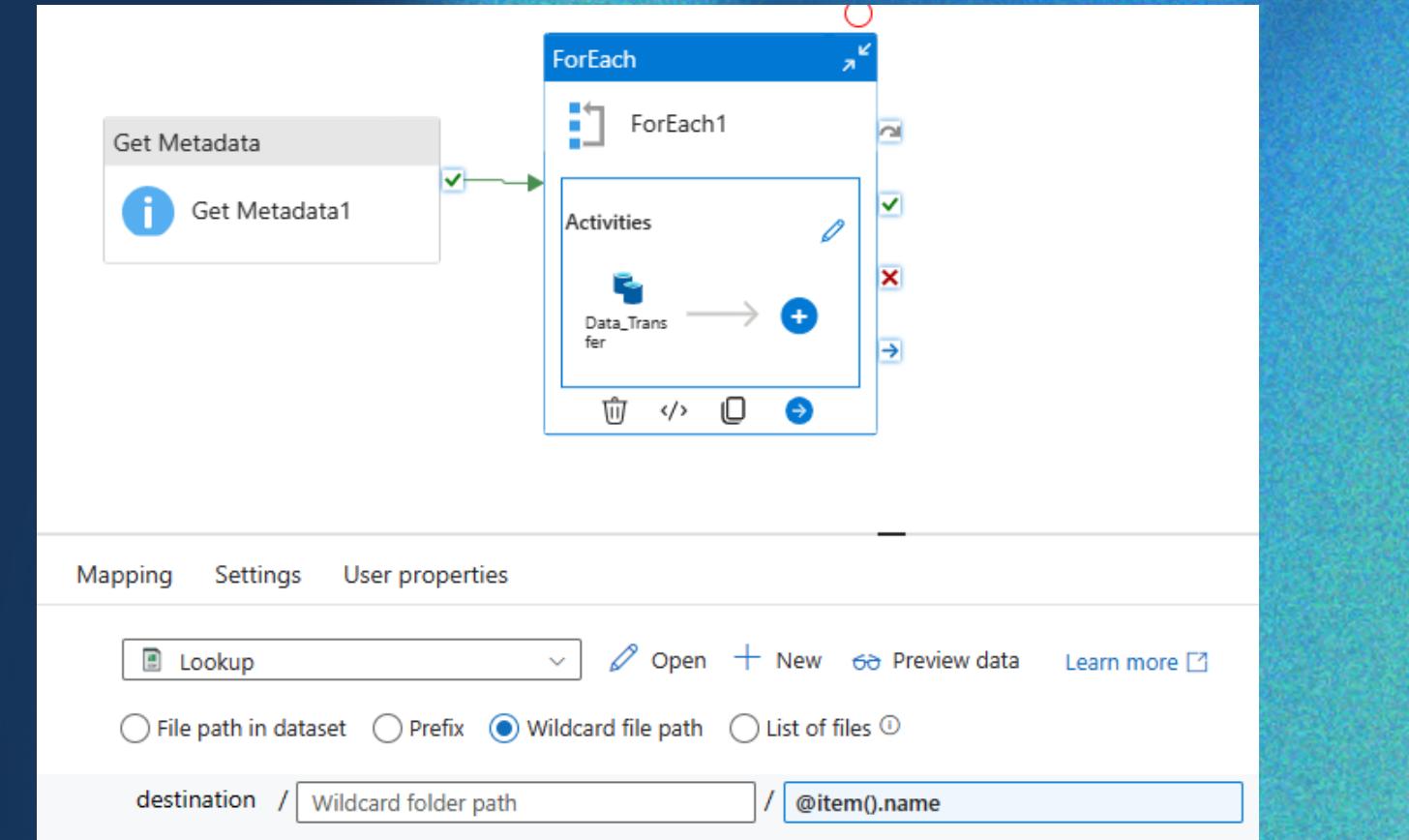
This study was effectuated in order to analyze and forecast recurring trends in breast cancer incidence and mortality rates

The data is obtained by the query tool which queries the data shared by Global Health Data Exchange provides

ETL

The data is extracted from the source storage account to the destination via the execution of the PowerShell script given by the commissioner.

After the transfer, an Azure Data Factory truncates all the previous tables and inserts the data auto-creating the tables if necessary



DWH

Once arrived in the database, the tables are processed by a stored procedure that unites the tables _1 and _2 into a third table.

The third table is going to be the main source of the dwh schema views presenting clean data.

```
12  SELECT
13    [Location]
14    , [Years] AS [year]
15    , cast(replace([SDI], '.', ',') as float) as SDI
16  FROM
17  (
18    SELECT
19      [Location]
20      , [1990], [1991], [1992], [1993], [1994], [1995], [1996], [1997], [1998]
21    FROM [stg].[region_sdi_total]
22  ) AS a
23  UNPIVOT (
24    [SDI] FOR [Years] IN (
25      [1990], [1991], [1992], [1993], [1994], [1995], [1996], [1997], [1998]
26    )
27  ) AS unpvt;
28 GO
```

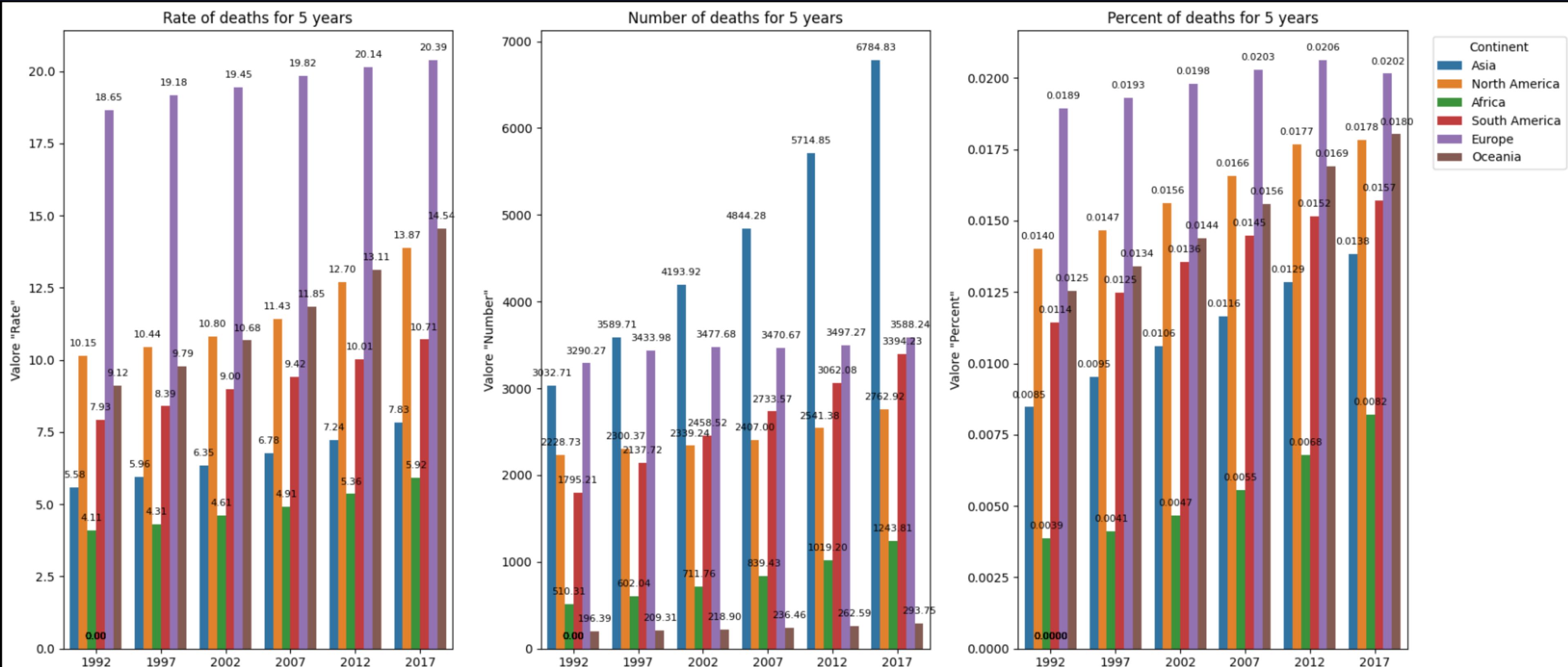
Country-Year-SDI unpivot

Supplementary Data

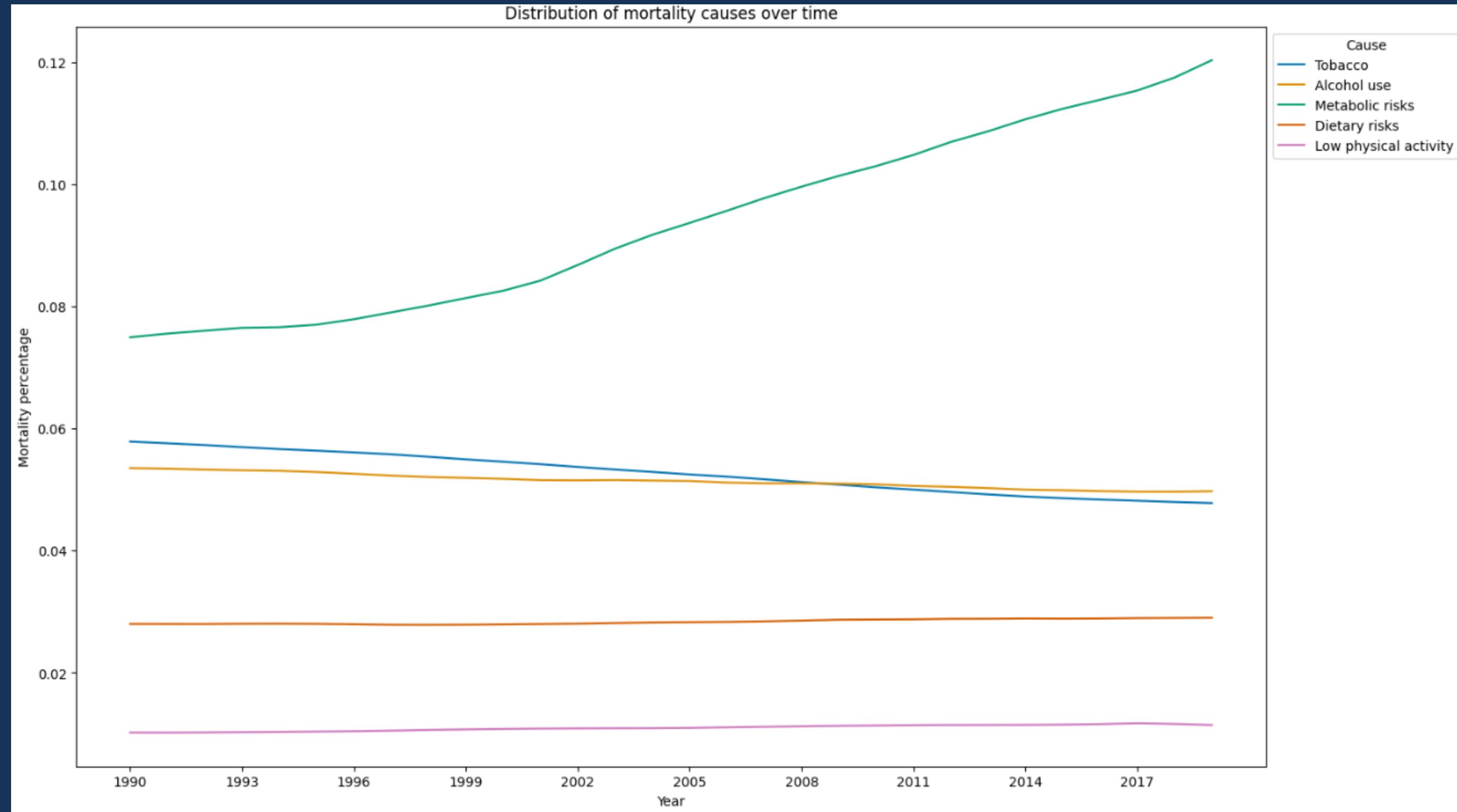
From the GBD query tool we extracted a table containing each country's sdi thorough the years.

EDA

Overview



Main cancer causes



DataFrame description

Code Markdown

```
1 stats_df = pd.DataFrame({df_name: df['val'].describe() for df_name, df in zip(['nation', 'region', 'region_sdi', 'percent'], [nation_df, region_df, region_sdi_df, percent_df])})
2
3 stats_df
```

✓ 0.0s Python

	nation	region	region_sdi	percent
count	18360.000000	630.000000	150.000000	1500.000000
mean	852.427615	9.676898	8.392783	0.263716
std	4540.150381	5.582889	4.605965	0.452682
min	0.000575	3.455839	3.694908	0.007054
25%	0.018210	4.830952	4.567857	0.033176
50%	6.541604	7.937454	6.352434	0.075622
75%	81.512116	15.054146	10.832264	0.230662
max	96306.310780	23.588512	16.859194	2.494516

Machine Learning

Here are the first rows of the tables used for machine learning:

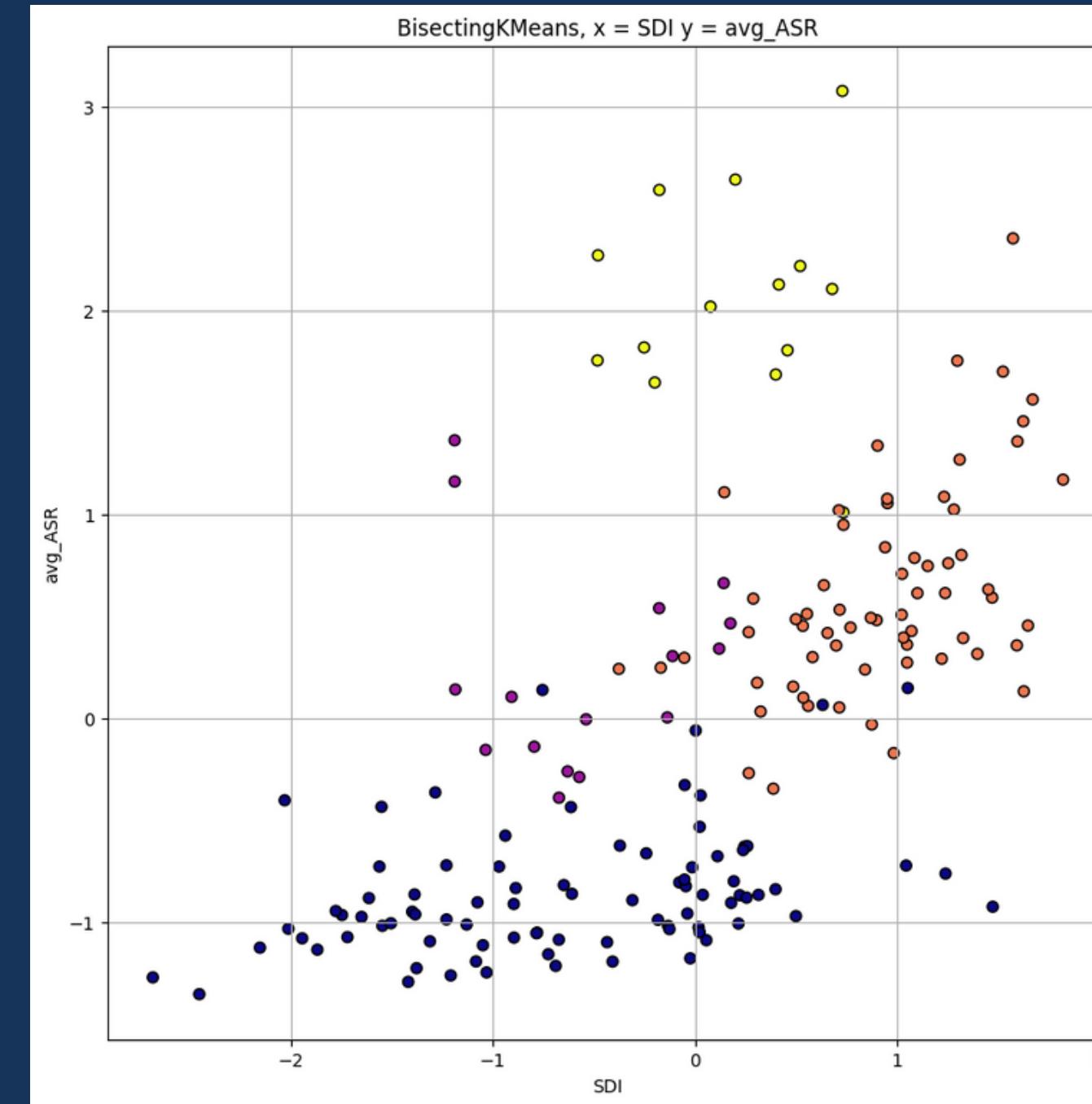
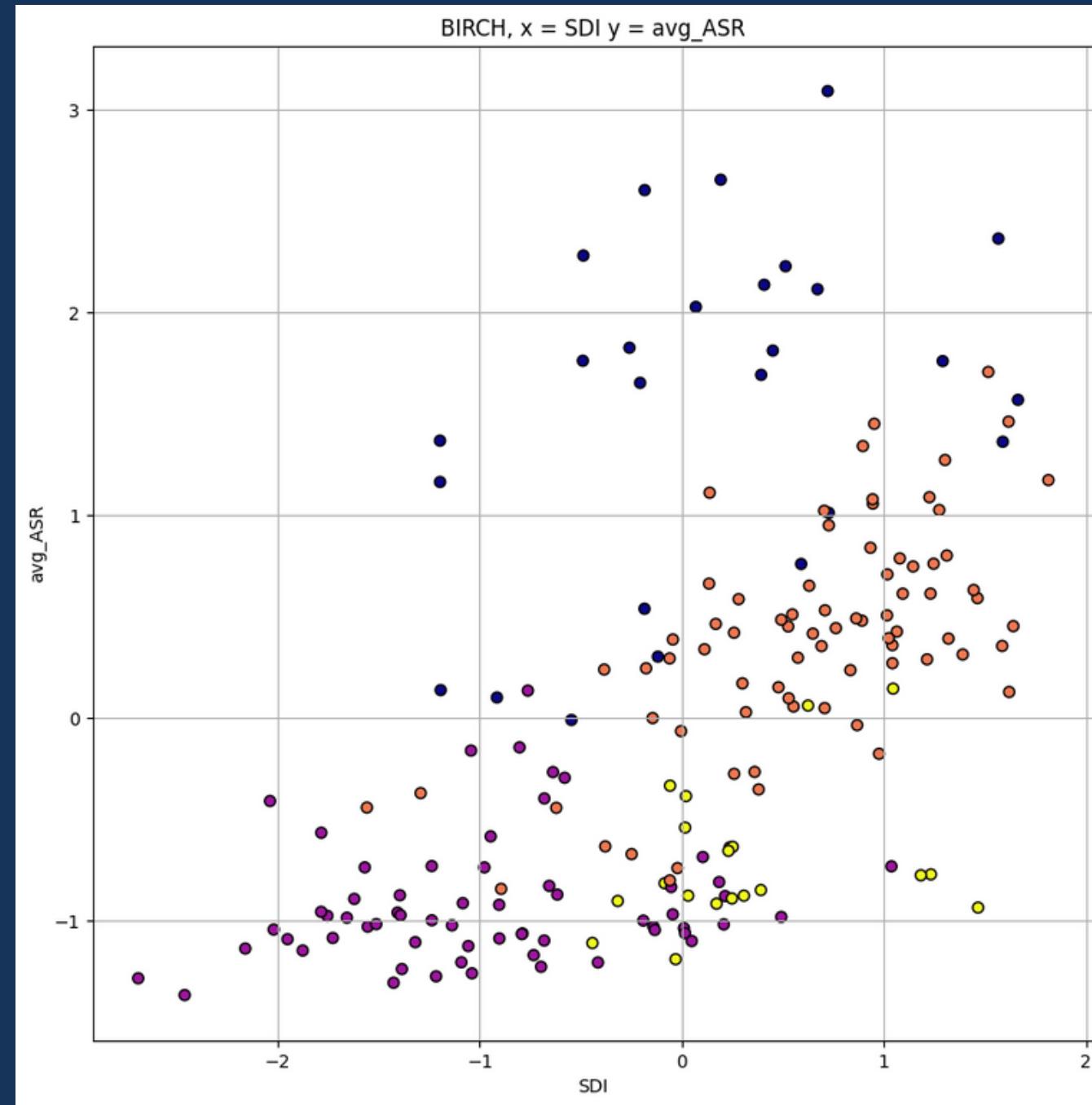
Clustering:

Afghanistan	0.2671425973 33336	0.2379000000 00002	10.1512055	6.1042191
Albania	0.2511383333 33337	0.6016333333 33328	6.9838551	5.2458868
Algeria	0.2699523053 33328	0.5547666666 66674	9.5287653	6.3349188
American Samoa	1.4833987699 9998	0.6583	16.3600687	11.9534951
Andorra	0.6863667633 33324	0.8634666666 66657	12.3316180	6.9802212

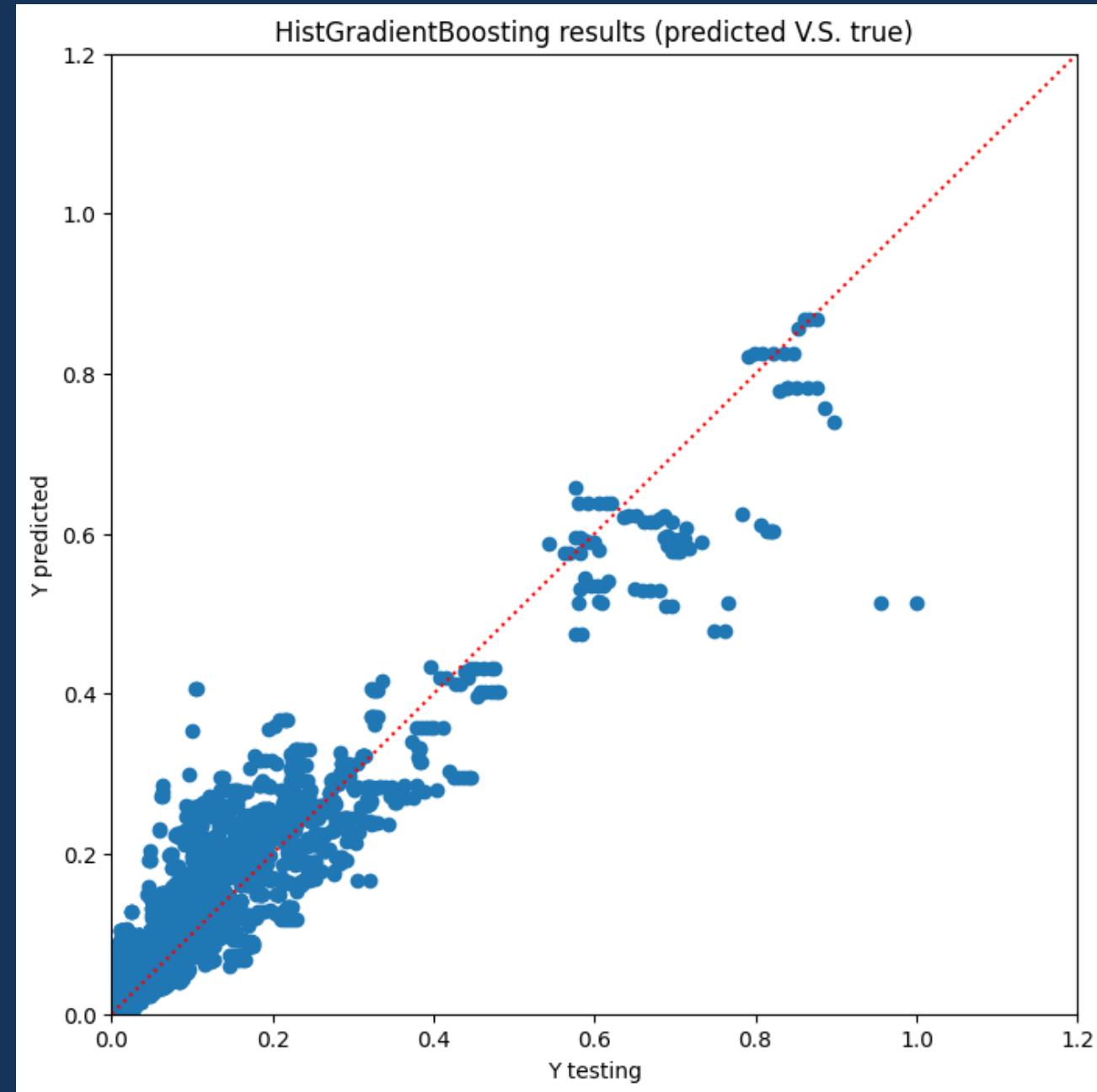
Regression:

Low-middle	Armenia	2019	Tobacco	0.5590176	0.689
Low-middle	Armenia	2019	Alcohol use	0.72227	0.689
Low-middle	Armenia	2019	Metabolic risks	1.8844075	0.689
Low-middle	Armenia	2019	Dietary risks	0.4724851	0.689
Low-middle	Armenia	2019	Low physical activity	0.1160354	0.689

Machine Learning - Clustering



Machine Learning - Regression



mae: 0.0275644619534122

mse: 0.0017321318536863658

r2: 0.8306138837447581

PowerBI



Thank you for your attention