Causes of Death

"What factors influence standardized death rates across European countries from 1994 to 2010?"

Team member: Uyen Pham, Linh Ha, Thuan Pham, Phuong Nguyen

Original Data

	DATAFLOW	LAST UPDATE	freq	unit	sex	age	icd10	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
0	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	F	TOTAL	A-R_V-Y	AL	2004	1267.1	NaN
1	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	F	TOTAL	A-R_V-Y	AT	1994	1124.2	NaN
2	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	F	TOTAL	A-R_V-Y	AT	1995	1105.8	NaN
3	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	F	TOTAL	A-R_V-Y	AT	1996	1096.1	NaN
4	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	F	TOTAL	A-R_V-Y	AT	1997	1060.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
384367	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	T	Y_LT65	Y10-Y34_Y872	UK	2006	3.7	NaN
384368	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	T	Y_LT65	Y10-Y34_Y872	UK	2007	3.8	NaN
384369	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	T	Y_LT65	Y10-Y34_Y872	UK	2008	3.8	NaN
384370	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	T	Y_LT65	Y10-Y34_Y872	UK	2009	3.6	NaN
384371	ESTAT:HLTH_CD_ASDR(1.0)	17/04/20 23:00:00	A	RT	T	Y_LT65	Y10-Y34_Y872	UK	2010	3.6	NaN

384372 rows × 11 columns

Project Process

IMPORTING LIBRARIES

13 libraries used for:

- Data processing
- Data visualization
- Machine learning models
- Label encoding

CLEANING DATA

1. Dropped & renamed columns
2. Removed redundant values (e.g., "TOTAL", "EU27_2020")
3. Filled 0.6% missing "death_rate" values with the mean

PREPROCESSING DATA

- Divide countries into 4 regions.
- One-hot encoding & label encoding for country code, sex, and age.
- Converted years into range (1-17).

ML MODELS

- Multiple Regression for its simplicity and interpretability.
- Random Forest Regression.

Cleaned Data

	sex	age	death_causes	country_code	year	death_rate	west_eu	east_eu	north_eu	south_eu	region
43935	0	0	ACC	AL	11	42.5	0	0	0	1	south_eu
43936	0	0	ACC	AT	1	93.2	1	0	0	0	west_eu
43937	0	0	ACC	AT	2	89.2	1	0	0	0	west_eu
43938	0	0	ACC	AT	3	93.7	1	0	0	0	west_eu
43939	0	0	ACC	AT	4	84.8	1	0	0	0	west_eu
...	...	...	...	...	...	...	...	...	...	...	...
247331	1	1	R	UK	13	3.3	1	0	0	0	west_eu
247332	1	1	R	UK	14	3.5	1	0	0	0	west_eu
247333	1	1	R	UK	15	3.4	1	0	0	0	west_eu
247334	1	1	R	UK	16	3.4	1	0	0	0	west_eu
247335	1	1	R	UK	17	3.3	1	0	0	0	west_eu

33600 rows × 11 columns

Creating Machine Learning Models

Multiple Regression Model (MRM)

Model Selection

- Dependent variable: death rate.
- Independent variables: death causes, age, sex, year, and 4 regions.

Key Metrics

- R-squared = 0.475, indicating the model explains 47.5% of the variance in death rate.
- P-values < 0.05 for almost variables, confirming their statistical significance.

OLS Regression Results

Dep. Variable:	death_rate	R-squared:	0.476
Model:	OLS	Adj. R-squared:	0.475
Method:	Least Squares	F-statistic:	1384.
Date:	Thu, 15 Aug 2024	Prob (F-statistic):	0.00
Time:	20:24:54	Log-Likelihood:	-2.5320e+05
No. Observations:	33600	AIC:	5.064e+05
Df Residuals:	33577	BIC:	5.066e+05
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	192.3463	9.324	20.628	0.000	174.070	210.622
C(death_causes)[T.A_B]	-54.1383	14.144	-3.828	0.000	-81.862	-26.415
C(death_causes)[T.C]	567.5975	14.144	40.129	0.000	539.874	595.321
C(death_causes)[T.D00-D48]	-65.2056	14.587	-4.470	0.000	-93.797	-36.614
C(death_causes)[T.E]	-11.0456	14.144	-0.781	0.435	-38.769	16.678
C(death_causes)[T.F]	-29.2474	14.162	-2.065	0.039	-57.005	-1.490
C(death_causes)[T.G_H]	-19.0346	14.144	-1.346	0.178	-46.758	8.689
C(death_causes)[T.I]	1528.3288	14.144	108.052	0.000	1500.605	1556.052
C(death_causes)[T.J]	174.3354	14.144	12.325	0.000	146.612	202.059
C(death_causes)[T.K]	26.1964	14.144	1.852	0.064	-1.527	53.920
C(death_causes)[T.L]	-73.6805	14.355	-5.133	0.000	-101.816	-45.545
C(death_causes)[T.M]	-69.3076	14.179	-4.888	0.000	-97.099	-41.516
C(death_causes)[T.N]	-30.2500	14.148	-2.138	0.033	-57.980	-2.520
C(death_causes)[T.O]	-21.8880	17.028	-1.285	0.199	-55.264	11.488
C(death_causes)[T.P]	-78.2733	14.162	-5.527	0.000	-106.031	-50.516
C(death_causes)[T.Q]	-77.2383	14.155	-5.457	0.000	-104.982	-49.494
C(death_causes)[T.R]	34.1130	14.144	2.412	0.016	6.389	61.837
age	-343.8730	4.948	-69.495	0.000	-353.571	-334.174
sex	76.9675	4.981	15.453	0.000	67.205	86.730
year	-3.1927	0.514	-6.208	0.000	-4.201	-2.185
west_eu	26.0523	4.595	5.670	0.000	17.046	35.059
south_eu	43.4323	5.028	8.639	0.000	33.578	53.287
east_eu	97.1529	4.725	20.561	0.000	87.892	106.414
north_eu	25.7088	5.639	4.559	0.000	14.656	36.762

Omnibus:	26101.453	Durbin-Watson:	0.063
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1518261.064
Skew:	3.257	Prob(JB):	0.00
Kurtosis:	35.281	Cond. No.	1.30e+16

Findings: Significant factors

- Death Causes:
  - Some causes (eg: Circulatory diseases) are the leading cause of death across European regions. This category includes heart diseases, stroke, and hypertension.
- Age:
  - Strong positive correlation with death rate: Older populations have higher mortality.
- Sex:
  - Males are at a higher risk of death compared to females.

Creating Machine Learning Models

Random Forest Regression (RFR)

Findings:

- R-squared = 0.945
  - The model explains approximately 95% of the variation in death rate (really strong level of fit).

- Difference between 2 models:
  - RFR: also captures complex, non-linear relationships.
  - MRM: assumes a linear relationship.

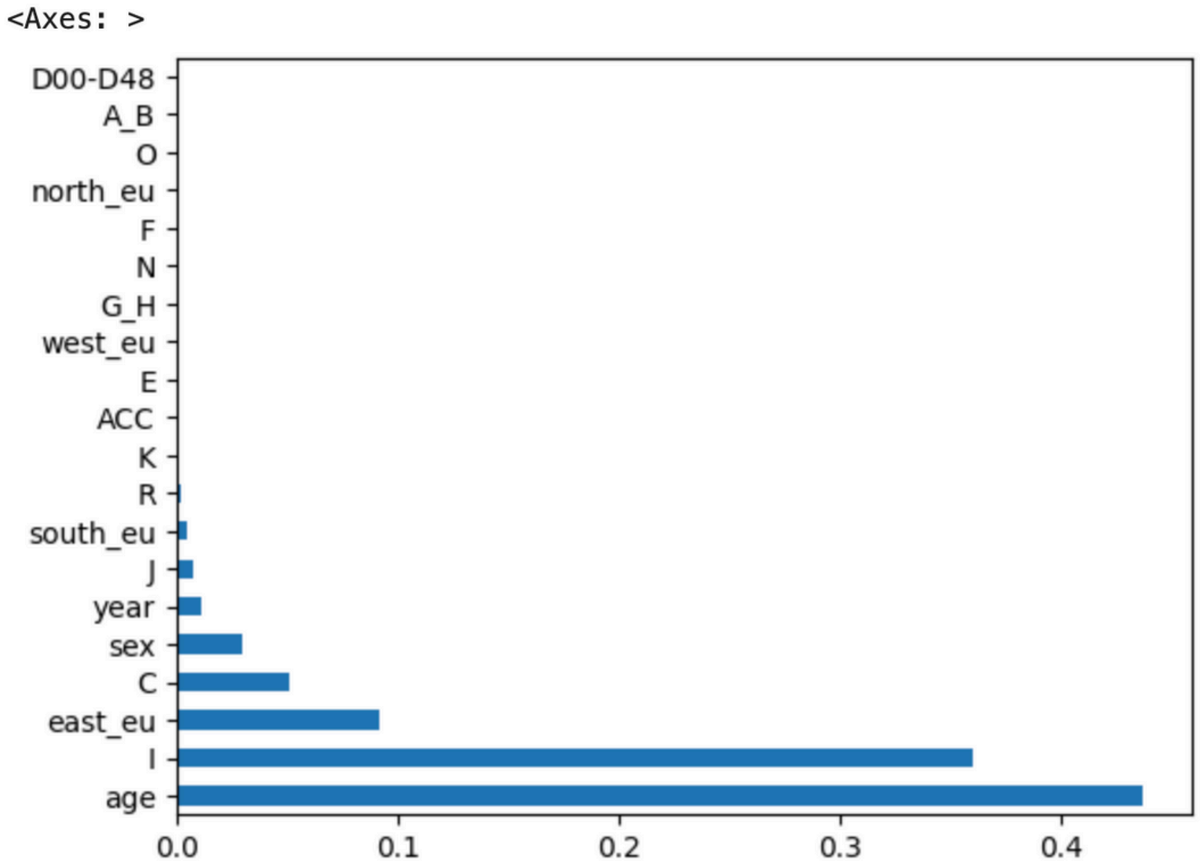


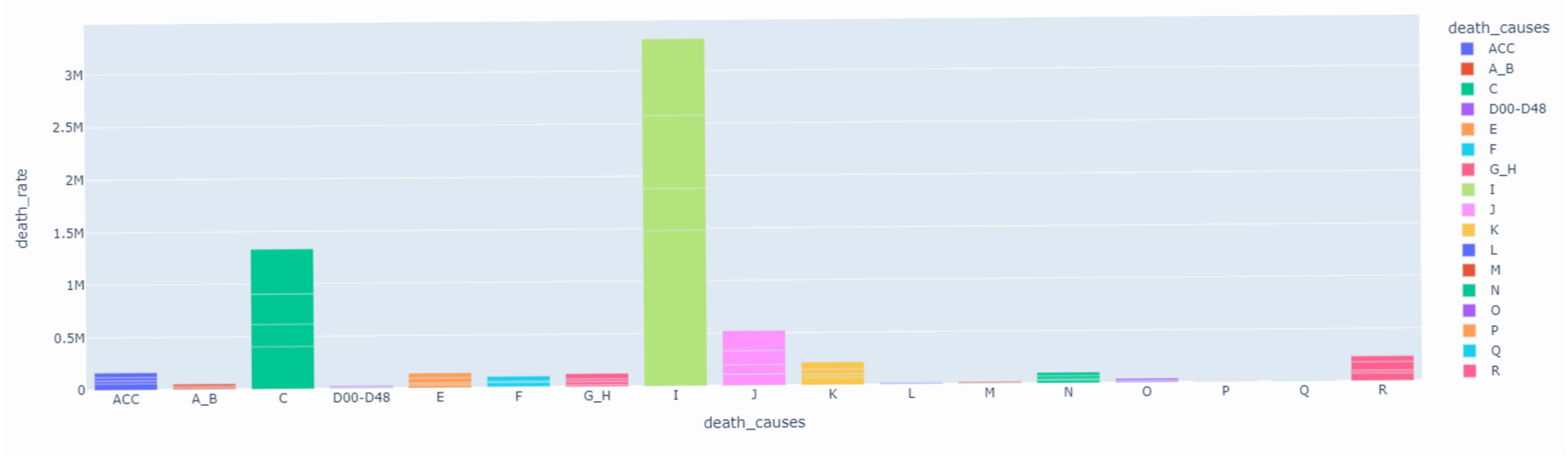
Figure: Feature Importance

Figure 1: Death Rate in European Regions (1994 -2010)

- Overall trend: EE & SE increased while NE & WEdecreased.
- NE had the lowest death rate, while EE had the highest.
- WE and SE showed intermediate trends.

=> Focus on healthcare improvements, especially in Southern Europe.

Figure 2: Total Death Rate of Each Death Cause in European Countries (1994-2010)



- "I" (circulatory system diseases) leads with the highest death rate, followed by C (malignant neoplasms, a type of cancer.)
- Other causes have significantly lower rates.

=> Prioritize interventions targeting circulatory diseases for impactful public health outcomes.

Policy Implications

- Focus on Major death causes:.
  - Launch public health campaigns on lifestyle changes (healthy diet, regular exercise, & smoking cessation).
  - Increase healthcare access to cardiovascular health services (regular screenings & early intervention programs).
  - Allocate resources towards medical research of circulatory diseases.

- Age-specific interventions:
  - Implement mandatory regular health check-ups for the elderly.
  - Expand the availability and quality of long-term care facilities.
  - Develop age-friendly environments that promote active aging (public transportation & accessible housing).

Data Source

https://ec.europa.eu/eurostat/databrowser/view/hlth_cd_asdr/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_hist