Exploring the relationship between socio-economic determinants and smoking rates in the EU

Big Data for Official Statistics

Santiago Vessi, 1958879

February 2025

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

- ▶ Introduction
- ▶ EDA
- ▶ Prediction, ANOVA & Random Forest
- ▶ K-means & PCA
- ▶ END

Introduction

- Smoking remains one of the leading preventable causes of disease and death worldwide, with significant social and economic implications
- The analysis leverages datasets from two reliable sources: **Eurostat**, the statistical office of the EU, and the **World Health Organization** (WHO)
- uncover patterns and correlations that may shed light on how social and economic disparities influence smoking behavior
- How smoking rates vary by income levels.
- Differences in smoking prevalence between genders and across age groups.
- The impact of educational attainment on smoking habits.
- The influence of urbanization, comparing smoking rates in urban versus rural areas.

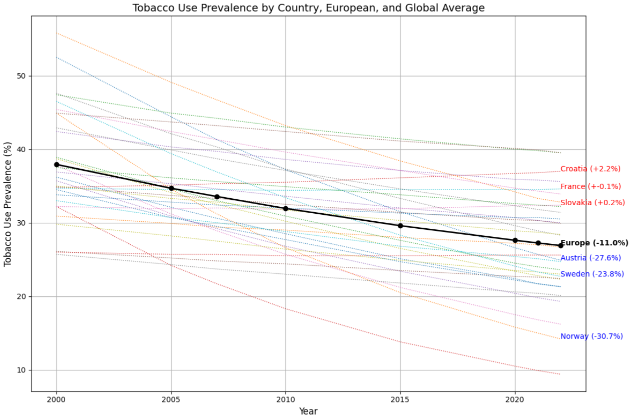
Table of Contents

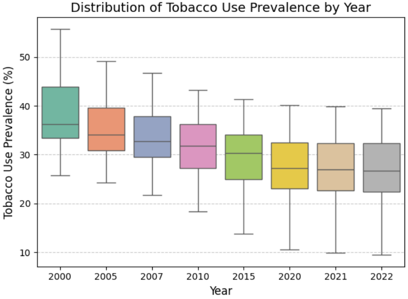
2 EDA

- Introduction
- EDA
- Prediction, ANOVA & Random Forest
- K-means & PCA
- END

WHO

2 EDA


Tobacco Use Prevalence by Country, European, and Global Average


Distribution of Tobacco Use Prevalence by Year

Relations between sex and tobacco consumption

Sex		2014	2019	Change
Females	Daily smoker	15.030	14.175	-0.855
	Non-smoker	80.485	81.596	1.111
	Occasional smoker	4.485	-4.221	0.264
Males	Daily smoker	24.664	23.046	-1.618
	Non-smoker	69.479	71.054	1.575
	Occasional smoker	5.855	5.918	0.063

Table: Difference between 2014 and 2019

Relations between age and tobacco consumption

Age		2014	2019	Change
From 15 to 24 years	Daily smoker	16.32	14.01	-2.31
	Occasional smoker	8.20	7.77	-0.43
From 25 to 34 years	Daily smoker	25.28	22.85	-2.43
	Occasional smoker	7.79	8.01	0.22
From 35 to 44 years	Daily smoker	24.15	23.57	-0.58
	Occasional smoker	5.89	6.15	0.26
From 45 to 64 years	Daily smoker	23.29	22.62	-0.67
	Occasional smoker	4.40	4.29	-0.11
65 years or over	Daily smoker	8.73	9.05	0.32
	Occasional smoker	1.79	1.98	0.19

Table: Difference between 2014 and 2019

Relations between education and tobacco consumption

2 EDA

Educ		2014	2019	Change
Levels 0-2	Daily smoker	20.37	19.09	-1.29
	Occasional smoker	4.12	3.98	-0.14
Levels 3-4	Daily smoker	23.50	22.13	-1.37
	Occasional smoker	5.70	5.49	-0.21
Levels 5-8	Daily smoker	13.81	13.01	-0.80
	Occasional smoker	5.96	5.70	-0.25

Table: Difference between 2014 and 2019

Relations between Income and tobacco consumption

Quintile		2014	2019	Change
First	Daily smoker	23.71	21.60	-2.11
	Occasional smoker	5.20	4.91	-0.29
Second	Daily smoker	20.22	19.18	-1.04
	Occasional smoker	4.48	4.58	0.10
Third	Daily smoker	19.65	18.58	-1.07
	Occasional smoker	4.83	4.58	0.10
Fourth	Daily smoker	19.20	18.38	-0.82
	Occasional smoker	5.28	5.18	-0.10
Fifth	Daily smoker	17.14	15.88	-1.26
	Occasional smoker	6.04	5.63	-0.41

Table: Difference between 2014 and 2019

Relations between urbanization and tobacco consumption

Urb. Level		2014	2019	Change
Cities	Daily Smoker	18.63	17.62	-1.01
	Occasional Smoker	5.51	5.82	0.31
Rural Areas	Daily Smoker	18.86	18.09	-0.77
	Occasional Smoker	4.57	4.53	-0.04
Towns and Suburbs	Daily Smoker	19.42	17.99	-1.43
	Occasional Smoker	4.87	5.04	0.17

Table: Difference between 2014 and 2019

Summary

- Both males and females show a decline in daily and occasional smoking, with males experiencing a larger drop in daily smoking (-1.618) compared to females (-0.855).
- Occasional smoking among males shows a small increase (+0.063), while it declines for females (-0.264).
- Significant decreases in daily smoking are observed among younger adults, particularly those aged 15–24 (-2.31) and 25–34 (-2.43), suggesting positive shifts in younger demographics.
- Older age groups (65 years or over) show a slight increase in both daily (+0.32) and occasional smoking (+0.19), indicating a potential rise in smoking habits among seniors.

Summary

- Daily smoking rates decline across all educational levels, with the largest reduction seen in group 2 (-1.37).
- Occasional smoking also decreases slightly across education levels, with no group showing an increase.
- Higher-income groups (fifth quintile) have a lower prevalence of smoking overall, with a decline in both daily (-1.26) and occasional smoking (-0.41).
- The largest reductions in daily smoking are observed in the first (lowest income, -2.11) and second quintiles (-1.04), highlighting progress among economically disadvantaged groups.
- Smoking decreases in all degrees of urbanization, with towns and suburbs showing the most significant decline in daily smoking (-1.43).
- Occasional smoking shows slight increases in cities (+0.31) and towns/suburbs (+0.1), while rural areas remain largely unchanged (-0.04).

Table of Contents

3 Prediction, ANOVA & Random Forest

▶ Introduction

▶ EDA

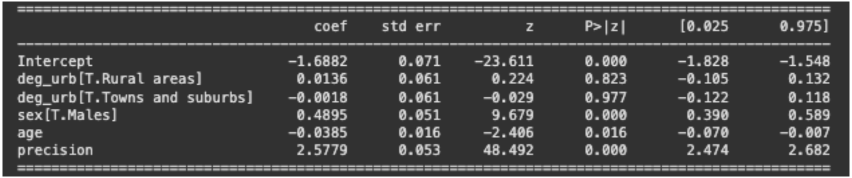
▶ Prediction, ANOVA & Random Forest

▶ K-means & PCA

▶ END

Beta Regression (Urbanization Levels)

- $OBS \sim age + deg\_urb + sex$
- R^2= 0.18
- MSE = 82.80

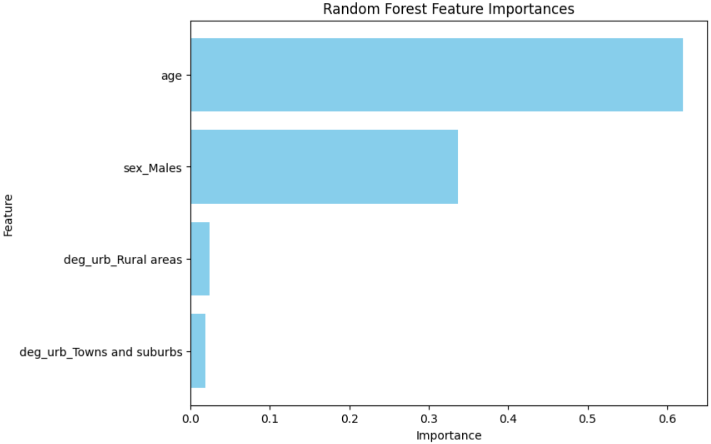


```
==============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                      -1.6882      0.071    -23.611      0.000      -1.828      -1.548
deg_urb[T.Rural areas]          0.0136      0.061      0.224      0.823      -0.105       0.132
deg_urb[T.Towns and suburbs]   -0.0018      0.061     -0.029      0.977      -0.122       0.118
sex[T.Males]                    0.4895      0.051      9.679      0.000       0.390       0.589
age                            -0.0385      0.016     -2.406      0.016      -0.070      -0.007
precision                       2.5779      0.053     48.492      0.000       2.474       2.682
==============================================================================================
```

ANOVA and Random Forest (Urbanization Levels)

	F	PR(>F)
deg_urb	0.008	0.905
sex	168.66	<0.001
age	11.14	<0.001

Table: ANOVA Results

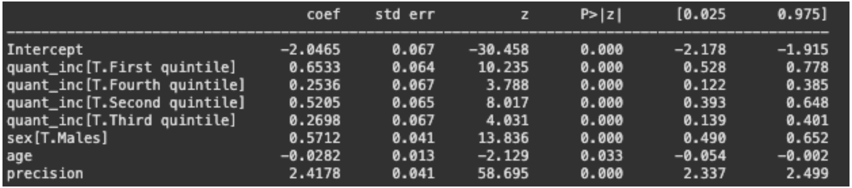


ANOVA and Random Forest (Urbanization Levels)

- Both analyses highlight **sex** and **age** as significant predictors.
- the Random Forest analysis emphasizes **age** more heavily than sex,
- while ANOVA suggests **sex** has a stronger statistical significance.
- **deg_urb** does not appear significant in either analysis, suggesting it has limited predictive power for this dataset.
- Random Forest Performance: the R^2 score suggests the model only partially explains the variance in the target variable.

Beta Regression (Income)

- $OBS \sim age + quant\_inc + sex$
- R^2= 0.08
- MSE = 113.89

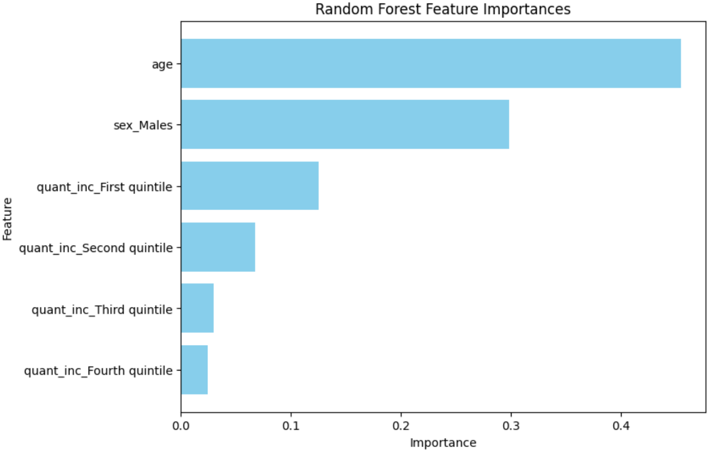


```
                                  coef    std err          z      P>|z|     [0.025     0.975]
--------------------------------------------------------------------------------------------
Intercept                      -2.0465      0.067    -30.458      0.000     -2.178     -1.915
quant_inc[T.First quintile]     0.6533      0.064     10.235      0.000      0.528      0.778
quant_inc[T.Fourth quintile]    0.2536      0.067      3.788      0.000      0.122      0.385
quant_inc[T.Second quintile]    0.5205      0.065      8.017      0.000      0.393      0.648
quant_inc[T.Third quintile]     0.2698      0.067      4.031      0.000      0.139      0.401
sex[T.Males]                    0.5712      0.041     13.836      0.000      0.490      0.652
age                            -0.0282      0.013     -2.129      0.033     -0.054     -0.002
precision                       2.4178      0.041     58.695      0.000      2.337      2.499
```

ANOVA and Random Forest (Income)

	F	PR(>F)
quant_inc	31.45	<0.001
sex	237.42	<0.001
age	14.05	<0.001

Table: ANOVA Results

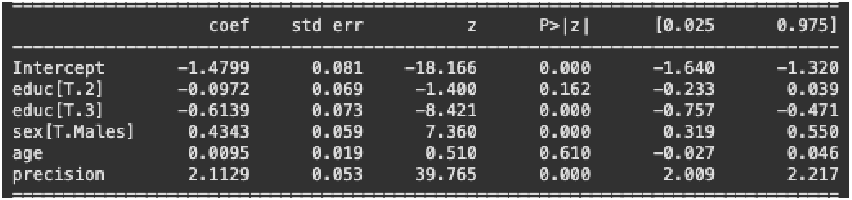


ANOVA and Random Forest (Income)

- Both analyses highlight the significance of **sex**, **age**, and **quant_inc** as predictors, but their relative importance differs:
- In ANOVA, **sex** and **quant_inc** dominate in terms of statistical significance.
- In Random Forest, **age** emerges as the most influential predictor, with **sex** as the second most important.
- The **first quintile** is most impactful, indicating a potential threshold or non-linear effect in how income affects the target variable.

Beta Regression (Education)

- $OBS \sim age + educ + sex$
- R^2= 0.24
- MSE = 134.88

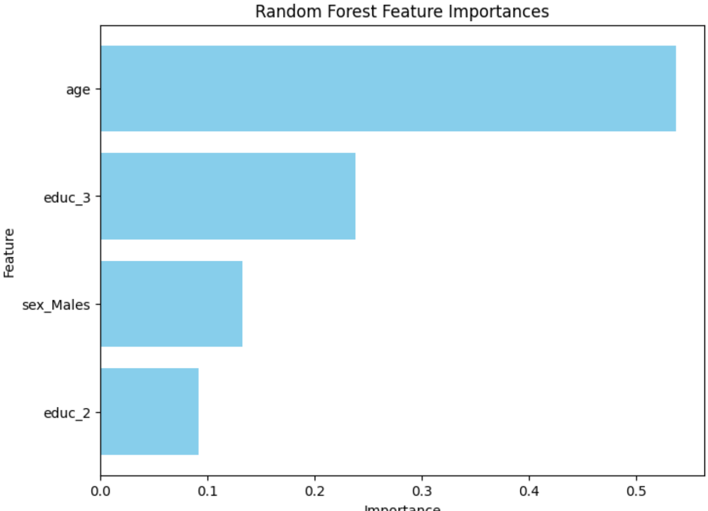


```
==================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        -1.4799      0.081    -18.166      0.000      -1.640      -1.320
educ[T.2]        -0.0972      0.069     -1.400      0.162      -0.233       0.039
educ[T.3]        -0.6139      0.073     -8.421      0.000      -0.757      -0.471
sex[T.Males]      0.4343      0.059      7.360      0.000       0.319       0.550
age               0.0095      0.019      0.510      0.610      -0.027       0.046
precision         2.1129      0.053     39.765      0.000       2.009       2.217
==================================================================================
```

ANOVA and Random Forest (Education)

	F	PR(>F)
educ	87.25	<0.001
sex	78.43	<0.001
age	7.78	<0.001

Table: ANOVA Results



ANOVA and Random Forest (Education)

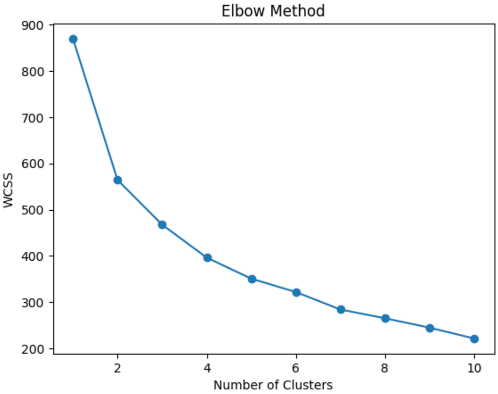
- ANOVA highlights **educ** and **sex** as the most statistically significant predictors, followed by **age** with a moderate effect.
- Random Forest elevates the importance of **age**, making it the most impactful feature, followed by specific education levels (**educ_3**)
- **Age** emerges as the most critical feature in Random Forest, despite being less impactful in ANOVA. This suggests age's relationship with the dependent variable might be non-linear or involve interactions that ANOVA doesn't capture

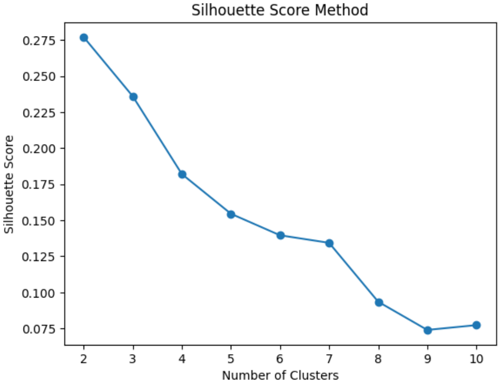
Table of Contents

4 K-means & PCA

- ▶ Introduction
- ▶ EDA
- ▶ Prediction, ANOVA & Random Forest
- ▶ K-means & PCA
- ▶ END

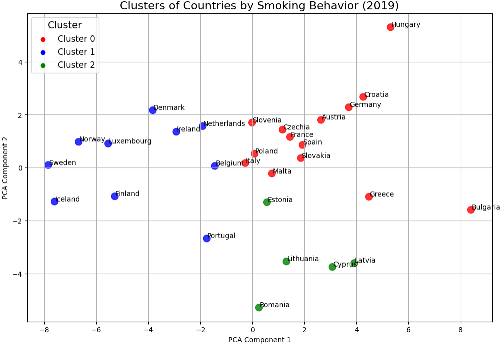
Clustering (Urbanization Levels)

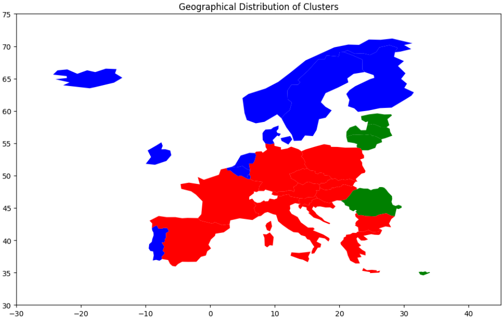




Clustering (Urbanization Levels)

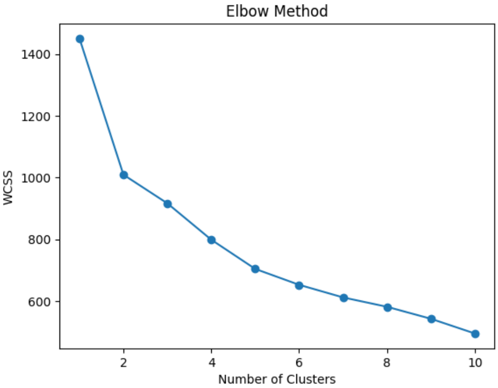
4 K-means & PCA

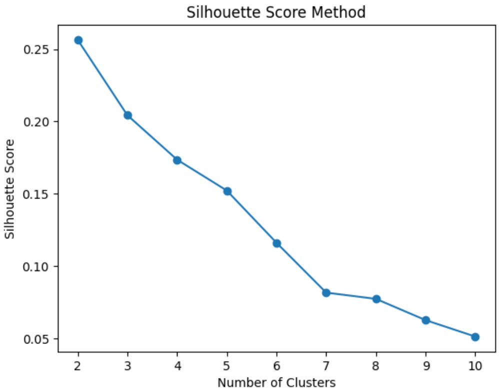




- **Cluster 0** (Rural/Non-Urban): Includes countries like Austria, Bulgaria, Croatia, and most Eastern European countries. These regions tend to have lower levels of urbanization, which might influence smoking habits through cultural or socioeconomic factors.
- **Cluster 1** (Highly Urbanized): Includes countries such as Belgium, Denmark, and Sweden, where higher urbanization levels are associated with different lifestyles, possibly including lower smoking rates due to public health measures or awareness.
- **Cluster 2** (Transitional Urbanization): Countries like Cyprus, Estonia, Latvia, and Lithuania fall into this cluster, representing areas in transition between rural and urban lifestyles.

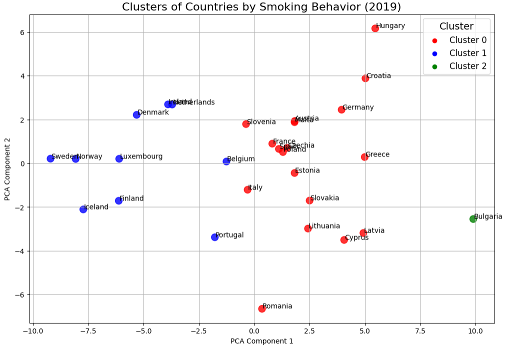
Clustering (Income)

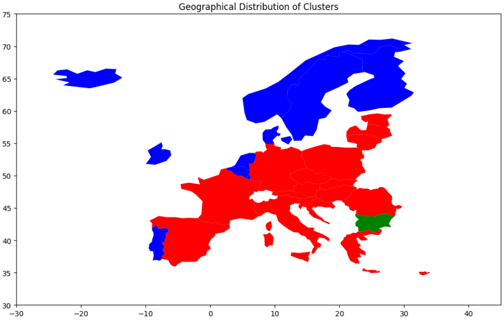




Clustering (Income)

4 K-means & PCA

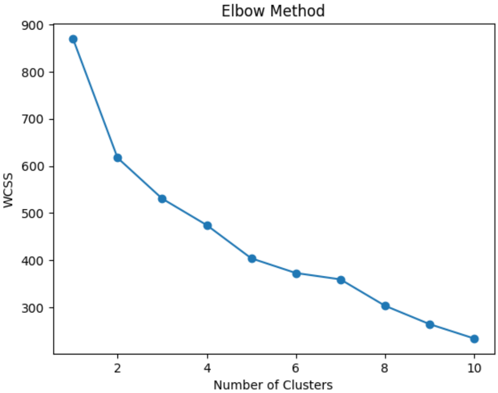


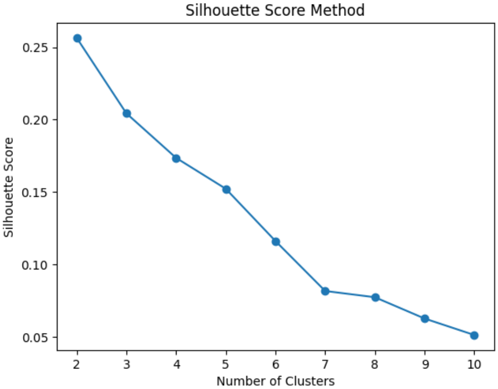


- **Cluster 0** (Low to Medium Income): Includes most countries like Austria, Croatia, and Poland. Despite being in similar income levels, these countries may have varying smoking habits due to other factors like cultural norms or public health initiatives
- **Cluster 1** (High Income): Countries such as Belgium, Denmark, and Luxembourg, where higher income levels might contribute to better health awareness and resources to reduce smoking.
- **Cluster 2** (Low Income): Bulgaria is the only country in this cluster, suggesting that economic constraints might influence smoking levels differently compared to other clusters.

Clustering (Education)

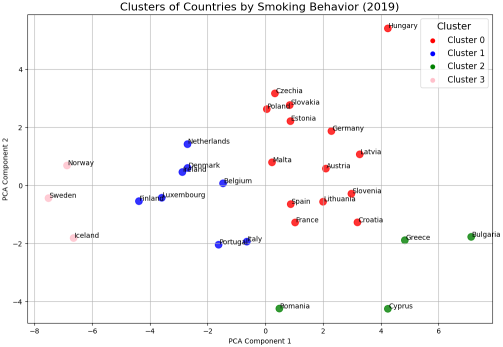
4 K-means & PCA

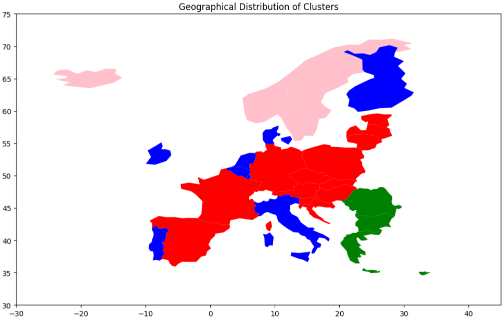




Clustering (Education)

4 K-means & PCA


Clusters of Countries by Smoking Behavior (2019)


Geographical Distribution of Clusters

- **Cluster 0** (Low Education): Includes countries like Austria, Bulgaria, and Slovenia. Lower education levels might correlate with higher smoking prevalence due to less awareness of the risks.
- **Cluster 1 & 2** (Medium Education/Specific Cultural) Medium education levels may indicate a balance where awareness and smoking rates are more stable. Also specific cultural or regional factors might influence smoking differently.
- **Cluster 3** (High Education): Iceland, Norway, and Sweden form this group, where high education levels likely contribute to reduced smoking rates due to better awareness and public health policies.

Table of Contents

5 END

- Introduction
- EDA
- Prediction, ANOVA & Random Forest
- K-means & PCA
- END