Project Report

Smoking Prevalence, Gender, and Youth Education across EU27 countries

Mathis Devoucoux, Justine Duplessis, Nils Ruetten

Introduction to Data Analysis with Python

December 20, 2025

Introduction

Our decision to investigate smoking behaviour through a gender and youth educational lens stems from a combination of personal observation and theoretical inquiry. As students living in Paris, we are regularly exposed to smoking within our social environment, prompting us to examine this phenomenon beyond personal experience. Our group chose to study smoking because it is a well-documented phenomenon that allows for clean, data-driven analysis. There is a thorough literature on the issue, which will be useful in guiding our hypotheses and conclusions. We are interested in the social patterns of smoking behaviour, particularly in how smoking prevalence may differ across gender and youth education in Europe. These aspects may offer some interesting insights into how social and cultural factors may shape health-related behaviour.

Using country-level data, we aim to identify whether consistent patterns of gender disparity emerge across Europe and how a factor such as youth education may interact with these trends. This approach should provide us with an understanding of smoking as a socio-cultural phenomenon and also offer some insights relevant for public health strategies tailored to the European context, where smoking remains a significant lifestyle and health issue.

Nevertheless, we are aware that our approach has limits. While we focus on gender and youth education as explanations for variation in smoking prevalence, there are likely other important factors that we ignore in our project. Furthermore, we rely on macro-level data that is furthermore heterogeneous in age groups when relating youth education to overall smoking prevalence. Our estimations and summary statistics will therefore not have the scope required to comprehensively explain smoking prevalence.

This framework leads us to formulate the two core research questions guiding our analysis. First, we seek to establish a baseline correlation by asking:

Research Question 1: Is there a significant difference in smoking prevalence between men and women in Europe?

Building on the findings from this univariate analysis, we then expand the scope to incorporate a key socioeconomic factor:

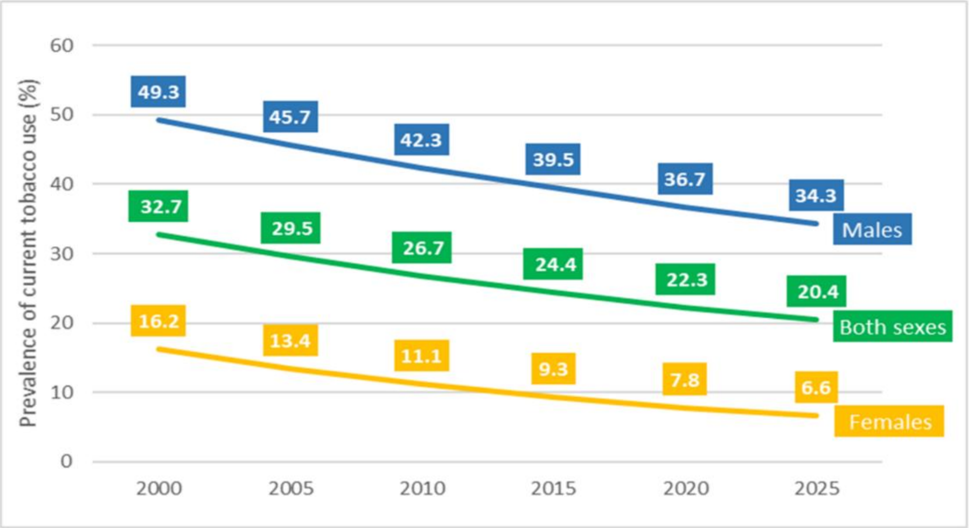
Research Question 2: Is smoking prevalence higher in European countries with lower youth education, as measured by the proportion of early school leavers (ESL)?

Literature Review

Smoking, a field of study in both public policy and academia

The issue of smoking and its effects is one of the major challenges of public health policies in the 21st century. The social perception of tobacco has evolved considerably, scientific studies have become common, and a general awareness of its harmful effects gradually emerged. The key question today is how welfare states should engage in the management and regulation of smoking. Should they tax cigarette packs, limit imports, or even prohibit tobacco consumption altogether? To what extent should the state intervene in regulation within liberal societies that uphold freedom as an ideal? Empirical studies indicate a general downward trend in global tobacco consumption (Figure 1).

Figure 1: *Global trends in prevalence of tobacco use among people aged 15 years and older, by sex*



WHO (2021). *Global report on trends in prevalence of tobacco use 2000-2025, fourth edition*. Geneva : World Health Organization

This positive development, however, is accompanied by a significant and concerning lag in mortality outcomes. As noted by the World Health Organization (WHO, 2021), the annual death toll attributable to tobacco „ can be expected to keep growing even once rates of tobacco use are in decline “a phenomenon explained by the fact that „ tobacco kills its users and people exposed to its emissions slowly” (WHO, 2021). This delayed effect is central to the seminal „cigarette epidemic model” developed by López, Collishaw, and Piha (1994), which describes the characteristic decoupling between peak smoking prevalence and peak mortality in populations, often spanning decades. Consequently, the persistent health burden of tobacco use remains a critical public health priority, even amidst declining consumption rates.

Our research is situated within this context and these recent debates about the means for the state to regulate without prohibiting tobacco consumption. The aim is to provide initial reflections on the groups most affected by smoking. The goal is to develop reflections that can, through quantitative analysis, inform public decision-making on public health issues. The state could then adapt its policy tools to target the specific social profiles of affected individuals.

There exists a plethora of studies, from both public policy and academic research, focusing on the effects of smoking and how it manifests in our societies. Within these, this project aims to adopt a social perspective, touching on the field of public policy regarding smoking, without necessarily providing operational suggestions. The study aims to improve the understanding of the phenomenon through the lens of two issues central to liberal societies during the 20th century: gender equality and the role of women, and education, particularly policies aimed at social "moyennisation" (Mendras & Duboys Fresney, 1994), i.e. the broadening of the middle class and democratizing access to higher education. In the following sections, we will successively analyse the various strands of literature that allow us to answer our two research questions, before concluding with the methodological references underpinning our quantitative analysis.

Gender and Smoking : Literature Review and Analytical Framework

The relationship between smoking and gender has undergone a significant historical evolution. While a pronounced male-dominated pattern created a persistent global gender gap for much of the 20th century (Waldron, 1991), subsequent decades have witnessed a notable convergence, often described as the „feminization" of smoking, driven by targeted marketing and shifting social norms (Amos & Haglund, 2000). Consequently, although men generally exhibit higher prevalence rates, the disparity has substantially narrowed in many high-income and European countries. Contemporary research reveals that this modern gender gap is not uniform but varies considerably across different European contexts, influenced by cultural and socioeconomic factors (Giskes et al., 2005), and that the determinants of initiation, addiction, and cessation remain distinctly gendered (Piper et al., 2010). This established yet dynamic landscape underscores the necessity of examining gender as a fundamental, but complex, social axis in understanding smoking behaviours.

Ritchie & Roser, 2024 provide a comprehensive quantitative analysis across multiple world regions, offering a broad perspective on the issue of smoking. Our objective here is to adopt a similar analytical approach using data sourced directly from the Eurostat database. The aim is to replicate this methodology while subsequently introducing an additional variable to elevate the analysis to a multivariate level, a point that will be elaborated in the following section.

Education as a Social Determinant of Smoking

The literature examining the link between smoking and educational attainment is less extensive than that focusing on income, which has been a primary lens for analysing addiction across social classes. Based on our academic perspective, we hypothesize that education level serves as a potent and distinct indicator for mapping smoking disparities, an intuition supported by the relative scarcity of focused studies on this variable. Huisman, Kunst & Mackenbach (2005) provide a foundational European analysis that is of particular interest to our inquiry. However, its reliance on data from the late 1990s represents a significant temporal limitation. Our study therefore aims to achieve two primary objectives: first, to update this research with contemporary data, reflecting recent socio-economic evolutions; and second, to deepen the analysis by conducting a multivariate examination that intersects gender with education level, directly addressing our second research question. This approach will allow us to offer a novel and more nuanced contribution to the field.

Methodological Framework and Analytical Tools

The quantitative analysis for this study will be conducted using the Python programming language, executed within the Visual Studio Code (VS Code) integrated development environment. This setup provides a robust and reproducible setting for data manipulation, statistical analysis, and visualization. We leverage core libraries, primarily pandas for data handling, numpy for numerical operations, and matplotlib as well as seaborn for generating graphs, to perform both our descriptive uni- and multivariate analyses.

This methodological approach is guided and informed by the instructional materials of the course.

Methodology

For our analysis we choose to merge three different Eurostat datasets because they provide aggregate information about the smoking prevalence as well as the educational level across European countries. We use this Eurostat data primarily, because it offers easy access and allows us to merge without great difficulty, as the unit of observation is similar and the merging keys are obvious.

We first conduct univariate analysis on the smoking dataset to obtain a simple mean difference analysis for smoking prevalence between men and women. The dataset used contains smoking prevalence data for population aged 15 or above in 2023 for all 27 EU countries. For this, we import the data, clean it and then run a simple regression of smoking prevalence on a gender dummy. We then aim to visualize the relationship of smoking and gender, by replicating the Our World in Data (2024) scatter, but with European countries in 2023. For each country a datapoint will indicate the smoking prevalence for men compared to women.

Afterwards, we aim to extend our questions scope and move from uni- to multivariate analysis. This will require importing an additional dataset and merging it with our smoking data. For the educational data we proxy youth education by the percentage of early leavers from education in 2023 separated for men and women. While this dataset includes data for more than the 27 EU countries, as it includes e.g. EEA as well, we restrict on countries for which smoking data is also available. Furthermore, as there is no female education data available for Luxembourg in 2023, we only consider the male percentage for this country. Then, we run a simple multivariate regression of smoking prevalence on both gender and youth education to clarify the influence of each variable on the smoking prevalence and put our result in a regression table as well as presenting a heat map of the correlations between the variables.

Results

Is there a significant difference in smoking habits between men and women in Europe?

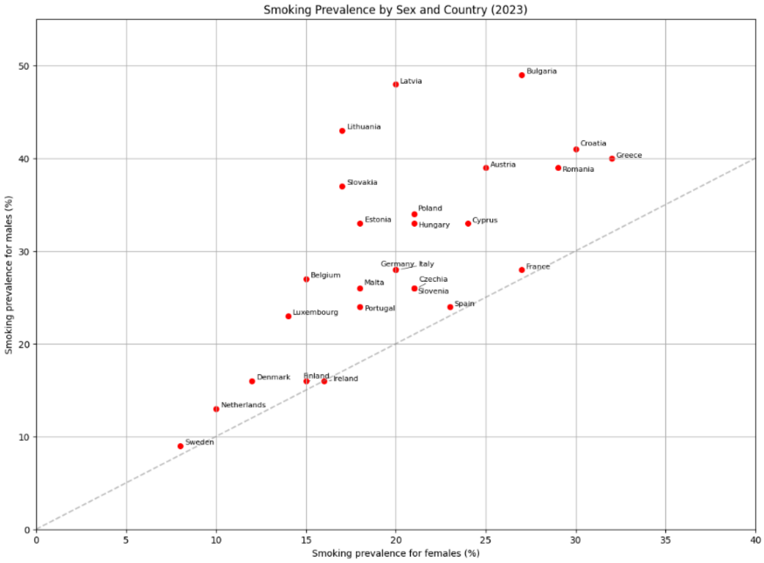
The result of our regression analysis via Python is given below (Table 1), although it has been truncated to keep only the main coefficients of interest. The intercept coefficient is 19.96, which means that the average smoking prevalence among women is 19.96% across the 27 EU countries. The coefficient on gender means that the smoking prevalence among men is 9.62% higher than for women, which means that, on average, 29.59% of men smoke. These results are consistent with other scholars' research.

Table 1*: Results of univariate regression of smoking prevalence on gender*

	Coefficient	Std. Error	t-statistic	p-value
Gender	9.63	1.46	6.59	0.00
Intercept	19.96	1.15	17.30	0.00

Additionally, we have produced a graph showing the smoking prevalence by sex for the year 2023 with the smoking prevalence for men on the y-axis and the smoking prevalence for women on the x-axis (Figure 2). As seen in the regression and the low p-value of the coefficient, the smoking difference by gender is highly significant. We can observe that all countries are above the 45 degree line, which confirms our previous finding: the smoking prevalence is higher for men than women in every country. This systematic observation indicates a deep-rooted habit linked to gender social norms.

Figure 2*: Smoking Prevalence by Gender and Country*



Own Visualization, modelled after OurWorldInData (2024)

Is smoking prevalence higher in European countries with lower levels of education?

The results of the multivariate regression (truncated) are below in Table 2. Our dependent variable is smoking prevalence. It has been regressed on the independent variables gender and percentage of early school leavers (ESL) as a proxy for youth education. The intercept coefficient is slightly higher than in the first regression (see Table 1): Here, the smoking prevalence among women is 20.621%, as it would be the prevalence for women in a country with 0% early school leavers. A 1 percentage point increase in early school leavers is associated with a decrease in smoking propensity by 0.06 percentage points, gender held constant, but this effect is indistinguishable from zero. We can therefore not find a relationship between our measure of youth education and smoking prevalence on the country level. At the same time, being a man continues to be associated with an increase in smoking prevalence at 9.55 percentage points for a constant level of youth education.

Table 2: *Results of multivariate regression of smoking prevalence on gender and ESL*

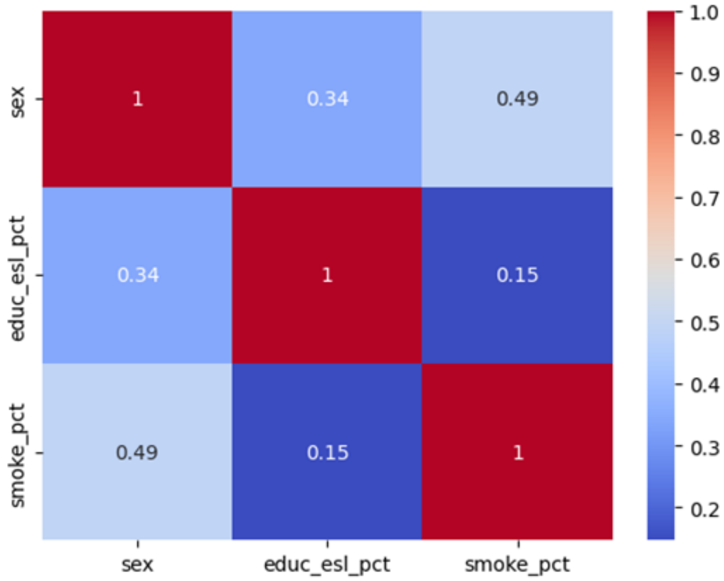
	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Gender	9.55	1.94	4.93	0.00
Early School Leavers (%)	-0.06	0.45	-0.13	0.89
Intercept	20.62	3.57	5.78	0.00

To further investigate the relationship between smoking prevalence, gender and early school leavers, we complement the multivariate regression analysis with a correlation heatmap (Figure X). This visualization presents pairwise correlation coefficients between the three variables across the EU27 countries (Figure 3).

The heatmap reveals a moderate positive correlation between gender and smoking prevalence (ρ = 0.49), confirming the descriptive and regression-based findings of a substantial gender gap in smoking behavior. Countries with a higher proportion of male smokers consistently exhibit higher overall smoking prevalence, reinforcing the central role of gender in explaining cross-country variation in smoking rates.

As seen before, the correlation between early school leavers and smoking prevalence is weak and positive (ρ = 0.15). This result is notable, as it does not align with the common expectation, derived from individual-level studies, that lower education is associated with higher smoking prevalence. At the aggregate country level, this weak correlation suggests that differences in educational attainment across EU countries explain little of the observed variation in smoking prevalence.

Figure 3: *Correlation heatmap of smoking prevalence, gender and early school leavers*



Own Visualization

The correlation between gender and education level is moderate (ρ = 0.34), indicating that being an early school leaver is more prevalent among men than women overall.

These insights from the heatmap are directly reflected in the regression results presented in Table 2. While the coefficient on education is negative, suggesting that higher educational attainment may be associated with lower smoking prevalence, it is not statistically significant. This lack of significance is consistent with the weak correlation observed in the heatmap and implies that, at the country level, education does not exert a robust independent influence on smoking prevalence once gender is controlled for. By contrast, the gender coefficient remains large and highly significant, confirming that gender differences are an important factor in explaining smoking prevalence in this dataset.

This highlights a key methodological insight: macro-level analyses may fail to capture social gradients that are well established in micro-level research, particularly when cultural, policy, and historical factors differ substantially across countries. Furthermore, our measure of youth education may be ill-suited for detecting country-level differences, as it assumes that todays percentage of early school leavers informs about youth education historically as well which shaped the smoking behaviour of earlier cohorts.

Conclusion

This project examined smoking prevalence across EU27 countries through the lenses of gender and youth education using recent Eurostat data. Our analysis provides evidence of a persistent and statistically significant gender gap in smoking behaviour: men exhibit substantially higher smoking prevalence than women in every EU country considered. This finding aligns with existing literature and confirms that gender remains a central axis structuring smoking behaviour in contemporary Europe.

By contrast, we find no statistically significant relationship between smoking prevalence and youth education as proxied by the share of early school leavers at the country level. While individual-level studies often document strong educational gradients in smoking, our results suggest that such patterns do not translate cleanly to aggregate cross-country comparisons. This discrepancy highlights the limitations of macro-level analysis and inferring individual behaviour from country-level data. Overall, our findings underscore the importance of gender as a key determinant of smoking prevalence in Europe, while also illustrating the methodological challenges of capturing education-related effects using aggregate indicators.

References

Amos, A., & Haglund, M. (2000). *From social taboo to "torch of freedom": The marketing of cigarettes to women*. Tobacco Control.

Giskes, K., Kunst, A. E., Benach, J., Borrell, C., Costa, G., Dahl, E., Dalstra, J. A., Federico, B., Helmert, U., Judge, K., Lahelma, E., Moussa, K., Ostergren, P. O., Platt, S., Prattala, R., Rasmussen, N. K., & Mackenbach, J. P. (2005). Trends in smoking behaviour between 1985 and 2000 in nine European countries by education. *Journal of Epidemiology and Community Health*.

Huisman, M., Kunst, A. E., & Mackenbach, J. P. (2005). Inequalities in the prevalence of smoking in the European Union: Comparing education and income. *Preventive Medicine*.

Lopez, A. D., Collishaw, N. E., & Piha, T. (1994). A descriptive model of the cigarette epidemic in developed countries. *Tobacco Control, 3*(3), 242–247.

Mendras, H., & Duboys Fresney, L. (1994). *La seconde révolution française: 1965–1984* (nouvelle éd. refondue et mise à jour). Gallimard.

Piper, M. E., Cook, J. W., Schlam, T. R., Jorenby, D. E., Smith, S. S., Bolt, D. M., & Loh, W. Y. (2010). Gender, race, and education differences in abstinence rates among participants in two randomized smoking cessation trials. *Nicotine & Tobacco Research*.

Ritchie, H., & Roser, M. (2024). Who smokes more, men or women? Our World in Data. https://ourworldindata.org/who-smokes-more-men-or-women

World Health Organization. (2021). *WHO global report on trends in prevalence of tobacco use 2000–2025* (4th ed.). World Health Organization. https://iris.who.int/handle/10665/348537 World Health Organization+1

Waldron, I. (1991). Patterns and causes of gender differences in smoking. *Social Science & Medicine*.

Additionally, our GitHub repository and the code for this report can be found here:

https://github.com/nilsr-spec/Python-Introduction-Smoking-Prevalence.git